学号 20153023600060

密级

武汉大学本科毕业论文

基于 Boosting 算法的
中国 A 股市场多因子投资策略研究

院（系）名 称：经济与管理学院

专 业 名 称 ：金融学

学 生 姓 名 ：陈梦玄

指 导 教 师 ：李斌 副教授

二〇一九年五月

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名：_____ 日期：_____

摘　要

机器学习算法运用在股票市场多因子模型上可以有效地捕捉因子之间复杂的相关关系，解决高维数据的问题。为了检验国外股票市场异象因子在中国 A 股市场的适用性，以及 Boosting 系列机器学习算法在金融投资学领域的适用性，本文基于中国 A 股市场量化异象因子数据库，运用一系列 Boosting 算法，包括极限随机森林（Extreme Random Forest）、梯度提升决策树（Gridant Boosting Regreesion）、极端梯度提升树（XGBoost）以及分布式高效提升器（lightBGM），对 96 个异象因子进行横截面超额收益回归预测。本文采用截面回归、时间窗口内模型集成（ensamble）的方式对股价进行逐月滚动预测，回测时间窗口分别采用 12 个月、36 个月、60 个月来进行检验。

实验结果表明，多因子模型表现优于单因子模型，Boosting 集成算法模型优于 Fama Macbeth 回归模型，更高级的 Boosting 算法优于较简单的 Boosting 算法。多因子模型 12 月多空组合（Long-Short Portfolio）相比的最优的前 5 个单因子模型年化收益率能提高 29.91%、夏普比率能提高 0.49；机器学习 Boosting 集成算法多空组合相比 Fama Macbeth 回归模型年化收益率能提高 10.04%、夏普比率能提高 0.55。

总体而言，时间窗口越长，模型所能接受的信号越多，各指标不显著的数目越少，表现越优异。除了 Fama Macbeth 回归和梯度提升决策树在时间窗口为 36 个月时的表现最好以外，其余模型均呈现时间窗口越长表现越优；时间窗口为 60 个月的分布式高效提升器模型夏普比率可达到 2.52、年化收益率可达到 64.63%。从运行效率上看，分布式高效提升器模型能提升计算效率 2.6 倍。在模型稳健性检验的部分，本文检验了市值加权多因子投资模型的绩效表现，结果表明简单加权组合构造的结果均优于市值加权。

相比 Fama Macbeth 回归，Boosting 模型能更好地识别因子有效性；多因子的因子筛选结果与简单的单因子的因子排序情况差异较大，说明多因子模型可以更好地识别因子之间的相关关系。交易摩擦类因子、动量因子、财务流动性因子在中国 A 股市场重要性排前三位，验证了其市场的有效性。

与国外市场相比，在中国 A 股市场中，各多因子模型之间的差距较小。模型预测的样本外的迪堡马里亚诺（Diebold-Mariano）检验结果表明所有模型均不存在显著差异。

关键词：股票横截面超额收益；Boosting 算法；机器学习；金融科技

ABSTRACT

Machine learning algorithm can effectively capture the complex correlation between factors and better handle multi-dimensional data in asset pricing models. This paper aims at checking not only the anomalies of stock market research in China's A share market but also the applicability of Boosting algorithms to stock investment. Based on *Quantitative Factor Database of China A-Share market,* this paper uses a series of Boosting algorithms, including Extremely Randomized Forest, Gridant Boosting Regreesion(GBRT), XGBoost and lightBGM, to predict the cross-sectional expected stock return with 96 factors. Cross-sectional regression is employed wih ensembling approach and the Long-Short portfolio is constructed by monthly rolling the time window. The Back-Test time window is 12 months, 36 months and 60 months respectively.

The results show that multifactor regression is better than the single factor regression; Boosting algorithm is prior to the Fama Macbeth regression which can increase the average annual yield by 29.91%, and Sharp Ratio by 0.49, compared with the average results of the Top 5 single factor models. The ensemble Boosting model can increase the average annual yield by 10.04%, and Sharp Ratio by 0.55 compared with Fama Macbeth multi-factor regression. The results of the advanced Boosting algorithm models performed generally better than those of the simpler Boosting algorithm models.

The longer the time window, the less the number of insignificant indicators, except for the Fama Macbeth regression and GBRT in 36 months; 60 months of lightBGM model performs best with the annual yield of 64.63% and Sharp Ratio of 2.52. In addition, the lightBGM model can effectively save 2-3 times the running time. In the robustness test, the performances of size-weighted multi-factor models are tested, and the results are not better than simple weighted ones.

Compared with Fama Macbeth Regression, Boosting models can better recognized the effective factors and the results of factor importance is disparate to that of single factors because the multi-dimensional paradox is released. The most important three categories are trading friction factors, momentum factors and financial liquidity factors, which proves that China's stock market process its endogenous effectiveness.

The differences between each model are smaller compared with American market. The out of sample Diebold-Mariano statistics show that all the models differ little.

Key words: Cross-Sectional Expected Stock Return; Boosting algorithm; Machine Learning; Fintech

目　录

4 结论与展望

1 绪论

1.1 选题背景

金融学的核心在于资产定价，而资产定价理论中最核心的思想可以归结为现值理论，即一项资产的市场价格应该等于其所有预期现金流的折现。在一个有效的市场上，如果市场达到均衡，预期收益率和未来现金流是同一个硬币的两面，即前者反映在分母上，后者反映在分子上。预期收益率包含无风险利率（资金的时间成本）以及风险溢价（对投资者承担风险的补偿）。风险溢价，即超额收益，其衡量是投资学、资产配置、组合管理中一个重要的话题。股票市场作为金融市场的一大组成部分，其风险溢价的衡量更成为了金融学领域最为广泛的研究内容。在股票市场投资的选股（picking）和择时（timing）两个方向上，如何从全市场范围内选择出错误定价的股票是学界和业界一直关注的热点，也即找到未被市场套利所消除的风险溢价。衡量风险溢价的目的不单纯在于寻找解释股票价格的因素，更重要是通过构造因子来实现对风险溢价的预测，进而实现对股票价格的预测。

在这个意义上，预期收益理论在股票投资的选股方面扮演了重要的角色。

预期收益理论（Antti Ilmanen,2011）[1] 可以划分为两个主要板块，一个是理性预期理论，另一个是考虑偏差的预期理论，后者更多的和行为金融学相联系。理性预期理论的开山鼻祖是资本资产定价模型（Capital Asset Pricing Model），该模型认为市场风险因子驱动着资产价格的变化，而且市场风险因子与股票的预期超额收益之间存在正向的线性关系，即在市场风险因子上的暴露越多，超额收益越大；它基于投资者理性、市场无摩擦、单投资期、一致预期等严格假设。到 20 世纪 90 年代，随着资本资产定价模型收到学术界的广泛认同，以及套利定价模型（Asset Pricing Theory）的引入，将风险因子从市场风险因子拓展至更广泛意义上的风险因子，除了市场风险以外的其他风险也被统一归为无法被市场所分散的系统性风险，比如说 Fama Franch 三因素（Fama Franch，1993）[2] 中的价值因子（净值市值比）和规模因子（市值）对风险溢价也存在较强的解释力度；Carhart（1997）[3] 提出了四因素模型；十多年后，Fama Franch 提出了五因素（Fama Franch，2015）[4] [5] [6]，增加了盈利性因子（profitability）和投资模式因子（investment pattern），因子投资理论受到学界的广泛认可。在此之后，多因子模型呈现爆炸式的发展，

横截面收益预测的因子越来越多，比如动量因子、反转因子、波动率因子、流动性因子、财务因子、盈利性因子、公司性质等等，甚至将行为金融学领域研究的投资者情绪因子、新闻事件影响因子等通过自然语言处理的方式引入到因子研究的框架中来，使得预期收益理论的两块内容的分界线越来越模糊，因子研究框架成为投资学领域最常用的研究范式。同时，因子的跨市场检验也成为了学界检验各个资本市场有效性的常用手段。

除了学术界，多因子模型也在业界也被广泛使用，但业界更多地使用短期技术指标作为因子，投资逻辑不够明晰，背后的经济学含义处于黑箱状况。反之，学界使用的因子从异象因子的角度出发，从解释股票预期收益的角度发掘其背后的经济学逻辑，其理论依据更加有力，能有效帮助投资者进行资产管理，帮助监管者评估公募基金和私募基金的风险敞口、识别并防范各类系统性风险。

除了因子本身的构造，多因子模型也对因子选股的数理实现提出了更高的挑战。在这一方面，近年来流行的机器学习算法被学者们关注并引入到多因子模型中，旨在更好地解决高维数据的问题，能帮助我们从海量的数据中抽取有效信息并进行识别，提高因子的识别效率，避免多因子模型因为因子数目的增加而降低单个因子所发挥的作用。

1.2 研究目的与意义

首先，本文研究基于中国 A 股市场量化异象因子数据库，是对于国外异象因子学术研究在中国市场的适用性的一次全面检验。受历史因素的影响，相比美国，中国的股票市场起步晚、发展欠成熟、产品单一、受控制力度大。受历史因素影响，两国股票市场存在很大的差异。所以在美国股票市场上有效的量化因子在中国未必有效，甚至存在有些因子作用完全相反的现象。本文的第一个目的就是检验国外文献中所提出的异象因子在中国 A 股市场的适用性，以此对以后中美股票市场对比研究产生一定的启示作用。综合国外文献中对异象因子的研究并检验其在中国市场的有效性，无论是对于国际股票市场差异与资产定价的进一步研究，还是对于业界从事全球资产配置的研究人士而言都具有不容小觑的意义。

其次，本文将机器学习算法应用在中国 A 股市场，也是对机器学习算法，尤其是 Boosting 系列机器学习算法在金融投资学领域适用性的一次检验与挑战。机器学习算法在物理化学生物等自然科学、医药地理等应用科学、心理学等社会科

学中均有应用，但是相对而言，其在金融学、经济学等社会科学中的运用还处于初步发展的阶段。相比于以往关于机器学习算法投资策略的论文，本文没有拘泥于一种或两种特定的策略，也没有泛泛地采取各个分支的机器学习算法策略，而是针对Boosting系列机器学习算法中主要的几个适用的分支算法来进行对比研究，使得结论对 Boosting 算法而言更有针对性。相较于其他机器学习算法，Boosting算法简单易于理解，且作为有监督学习的算法在众多领域已经取得了较好的预测效果，本文倾向于探究其在金融投资领域的效果。

本次论文作者与其研究团队[1]共同构造中国 A 股市场量化异象因子数据库，对单因子进行数据收集、整理，对单因子有效性进行显著性检验，对因子进行分类；针对多因子部分，本文作者通过机器学习量化多因子模型预测结果来构建投资组合，并与 Fama Macbeth 回归（Fama Macbeth，1973）[7] 进行对比，通过样本外滚动回测的实现来对模型进行检验。本文从理论追溯、技术手段、方法实现等角度深入探究机器学习算法在金融领域的应用，为未来金融科技的深入研究应用打下基础。

[1] 数据库研究团队组长为武汉大学经济与管理学院李斌副教授，组员包括武汉大学经济与管理学院 2018 级研究生邵新月、武汉大学经济与管理学院 2018 级研究生李玥阳、武汉大学经济与管理学院 2015 级本科生岳阳和本文作者。

1.3 多因子策略及量化研究文献综述

1.3.1 多因子量化投资相关文献综述

对资产的差异与收益之间的关系进行解释是金融的基本问题之一，这一问题可以归结为资产定价问题。而资产定价的研究基本分为时间序列上的预测和截面数据上的预测。

在股票截面数据研究上，Fama, French（1992）[15] 基于之前的研究首先用 10 个因子来预测收益，但是没有解释多个因子之间的潜在相关性。随后，多因子的研究运用在某一类量化因子情绪指标的构建中，比如与股票盈利性、市场情绪、交易成本有关的指标。比如，Stambaugh, Yu 和 Yuan（2012）[16] 用 11 个与投资者情绪有关的异象指标解释了风险溢价。Novy-Marx 和 Velikov（2015）[17] 用 23 个异象指标解释了考虑交易成本的投资组合的表现。

Lewellen（2013）[18] 构建了 15 个公司基本面因子的 Fama Macbeth 回归，并且发现 12 个月为最佳的滚动回归窗口，这一时间窗口的长度被后来的研究所广泛应用。Green. etal（2013）[19] 用 93 个因子构建多因子模型，并且发现当 t 统计量的绝对值大于 3 时，因子显著性的稳定情况更佳。理论共识得到进一步发展，多因子研究趋于成熟。

近年以来，随着单因子以及具有某一类特征的因子论文的不断累积，因子数目得到了迅速的拓展，为多因子模型的发展奠定了理论和数据上的基础。McLean 和 Pontiff（2016）[20] 用 97 个异象指标研究了上市后收益的盈利性，并通过样本区间前和样本区间后的检验来看学术因子研究中是否存在时期选择等数据挖掘的问题。Green. etal（2017）[19] 还用 40 年美国市场的近 330 因子进行多因子超额收益预测模型的构建。Hou, Xue 和 Zhang（2017）[21] 复制了 447 个异象因子指标，在剔除小盘股并使用等市值法时，仅 85%共 286 个异象因子在 5%的置信水平下保持单因子显著，仅 115 个异象因子在多因子模型中显著，这一方面说明在因子论文中普遍存在样本挑选和使用测试流程使得结果表现更优的问题，另一方面说明资本市场比过去认为的更有效，至少因子理论能有效地解释它。

在因子研究领域，中国 A 股市场因子也因为中国股票市场的迅速发展和中国股票市场独有的特点（Hu. etal，2018）[22] 而广受国内外量化投资学者的关注。潘莉、徐建国（2011）[23] 探索构建适用于中国 A 股市场的三因子模型，发现市盈率

和市值特别显著，市值背后既有风险因素又有特征因素，而市盈率主要表现为特征因子的特点。李志冰等（2017）[24] 考察五因子模型在中国股市不同时期的应用，并在中国市场上对比了 CAPM、三因子、Carhart 四因子、五因子，发现股改扭转了盈利能力、投资风格、动量因子这三个因子的效果。Guo. etal（2017）[25] 验证了 Fama 五因子模型在中国市场上的表现情况，发现盈利性因子能有效提高股票的边际收益率，投资性因子基本上“冗余”。Jason. etal 在 2017 年的工作论文[26] 中对比研究了美国股票市场和中国 A 股市场中，以探究在市场制度、金融报告准则、市场微观结构以及投资者行为不同的情况下，因子投资是否存在区别。结果发现，因子投资策略在中国确实有效：通过因子投资组合的构造，中国 A 股市场能取得比原来更加有效的投资策略。

与 Hou, Xue 和 Zhang（2017）[21] 的研究类似，清华大学五道口金融学院 2018 年发布的《中国 A 股市场量化因子白皮书》[27] 系统地研究了中国 A 股超额收益的决定性因素，根据 A 股股票交易数据和财务报表数据构建了 56 个量化因子，并逐一进行了单因子检验，发现在 1997 年 1 月至 2017 年 12 月期间，仅有 13 个有效因子。朱英伦、刘杰（2018）[28] 总结并检验了近年来在中国金融、会计、经济和管理等期刊上发表的因子文章，系统性地阐述了中国股票市场因子的有效性，发现在中国市场有近 9.67%（29/300）在美国最早发现的因子存在超额收益，并把这些有效的因子分为市场类因子、基本面因子、估值类因子和事件类因子四大类。

1.3.2 机器学习量化投资相关文献综述

随着因子数目的增多，简单的多因子模型带来的计量上的问题需要更高级的方法来解决。McLean 和 Pontiff（2016）[20] 用 93 个因子研究了多因子模型中的多重共线性问题，发现有 26 个因子在 Fama Macbeth 回归中 t 的绝对值大于 3 的情况下有多重共线性。Green. etal（2017）[19] 还用 40 年美国市场的近 330 因子进行多因子超额收益预测模型的构建，发现正交化处理之后的因子比随机选择的因子盈利概率更高。McLean 和 Pontiff（2016）[20] 和 Green. etal（2017）[19] 的研究都说明了现有的多因子研究中可能存在人为的数据挖掘问题。多因子模型中的多重共线性、高维数据、数据挖掘等的问题需要更高级的量化手段来解决。

于是，在方法的创新上，机器学习算法也逐渐被引入量化因子投资研究领域。Moritz.etal（2016）[29] 引入了一种新的因子投资组合构造方法，即基于树的条件投

资组合筛选方式（Tree-based Conditional Portfolio Sorts），其绩效表现优于传统的 Fama Macbeth 回归。Serhiy（2017）[30] 改良了传统的三因子、四因子、五因子，用广义线性回归（GMM）的方式对股票横截面收益进行预测。谢合亮、胡迪（2017）[31] 引入 LASSO 和弹性网（Elastic Net）两类方法进行因子筛选并确定因子权重最后构造投资组合。Serhiy, Stefan, Shrihari（2018）[29] 研究了主成分分析的因子模型作为缩减的因子模型在股票横截面收益预测中的理论推导与应用，以此来探究特征和方差在模型中是否可以被区分开来讨论。

机器学习算法应用在股票收益预测领域, 除了直接作为预测模型，另一个发展方向是作为特征或者股票筛选方式。在特征筛选上，Tsai, Hsiao（2010）[33] 使用主成分分析（PCA）、遗传算法（GA）、决策树模型（CART）来进行特征筛选，并将三个模型混合起来，在股票预测模型中可以过滤 80%不具有代表性的特征。在股票筛选上，Feng, Polson, Xu（2018）[34] 尝试使用深度学习的方法使用高维公司特征来产生多空组合的因子。他们使用非线性的方法先对股票进行筛选，通过隐神经网络算法来最小化股票收益的 Alpha。结果表明，神经网络算法确实可以提高某些异象因子的显著性。

在机器学习算法的横向对比上，Gu, Kelly, Xiu（2018）[35] 将多种机器学习方法结合并对比，包括广义线性模型、降维方法、增强回归树、随机森林、神经网络；并提出使用样本外收益预测的 R^2 作为模型评价标准；结果表明树和神经网络的方法表现最好，在因子中动量、流动性和波动性因子表现最好。Gu, Kelly, Xiu（2018）[35]使用了样本外 R^2 以及样本外的迪堡马里亚诺统计量（Diebold, Mariano, 1995）[36] 作为评价多个机器学习预测模型好坏的指标。在此之前，Bryan, Seth（2013）[37] 首次使用样本外预测的 R^2 作为评价指标，横截面上的账面市值比因子的截面预测回归样本外预测的 R^2 达到 0.9%，证明用现金流因子预测股票收益具有鲁棒性（robustness）。

Ondrej, Martin,（2018）[38] 使用历史文献中的 153 个股票市场异象因子来合成因子，结果发现用机器学习方法合成因子比单因子的可获得的收益率高四倍，夏普比率比传统的投资组合构造方法高两倍。同时，他们在国际股票市场上进行了更加广泛的检验，结果发现，除了美国股票市场，世界其他股票市场投资组合的样本外表现都不如美国市场，而美国股票市场的实证研究使用的因子可以捕捉到

其他股票市场的大部分超额收益。

除了将机器学习算法应用在美国股票市场的研究，中国股票市场上机器学习算法也得到了国内外学者的注意。中国股票市场经历了快速的成长，成为了世界第二大股票市场，也是全球资本市场的重要组成部分。理解中国股票市场定价的潜在规律对于全球投资者、学者以及监管机构来说越来越重要。Jiang, Tang 和 Zhou 的论文《公司特征与中国股票》（2018）[39] 是将机器学习和因子模型结合在一起的国内首创性研究。他们用 75 种公司特征预测中国股票跨行业的市场收益。除了使用传统的 Fama-Macbeth 回归，还使用了大数据或者机器学习的算法，如主成分分析（PCA）、偏最小二乘法（PLS）、预测组合（Forecast Combination）。75 个公司特征均来自于历史文献，并被划分为六个类别：价值与成长性、投资性、盈利性、动量、交易摩擦和无形资产类。结果表明，偏最小二乘法在中国股票市场上因子模型的表现最好；同时，与交易摩擦、动量、盈利性相关的指标在中国市场对股票收益预测性最好。

目前因子研究的发展趋势一是扩展历史文献中研究构建过的单因子，在中国市场中检验截面因子的有效性，二是运用机器学习算法在多因子回归中识别出对中国股票市场有效的因子。本次毕业论文研究正是顺着这两方面的发展趋势来进行的。

1.4 研究内容与方法

首先，本文将对国内外各因子投资的相关文献的收集、整理、分类，并进行初步研读，详细了解相关因子的计算逻辑和经济学内涵，整理成 Excel 表格，结合因子本身和历史文献的分类对各个因子进行系统性的细致的再分类，为实证研究中数据收集整理打下基础。

本文数据的获取方式主要是从 Wind 数据库获取股票交易数据，从 CSMAR（国泰安数据库）获取财务报表数据，根据国外量化异象因子研究来计算各个因子指标，并构建中国 A 股市场量化异象因子数据库。本次毕业论文的数据的时间区间为 1996 年 12 月-2018 年 10 月。股票池为全部 A 股，并对股票池进行了细致的处理，如剔除金融股（按照证监会 2012 年行业分类），ST 以及 ST*股，剔除所有股票停牌期间数据，剔除所有股票上市一年内的数据。数据频率为股票月度数据，对于数据频率比月度频率更加稀疏的数据，如季度公布的财务数据指标，为了避免使用未来数据，采取填充最近会计期间内可得的数据进行填充。然后，对每一个月度因子收益率数据进行描述性统计分析，计算单因子投资组合的年化收益率、累计收益率、波动率、最大回撤、夏普比率，画图分析单因子投资组合净值随时间变化的情况，并对其收益率的显著性进行检验。

其次，本文构造一系列多因子投资策略模型，包括极限随机森林（Extreme Random Forest）、梯度提升决策树（Gridant Boosting Regreesion）、极端梯度提升树（XGBoost）以及分布式高效提升器（lightBGM），根据 96 个异象因子进行横截面超额收益回归预测模型来构造多因子投资策略模型。通过 12 个月、36 个月、60 个月，这三种时间窗口滚动来构建多空组合（Long-Short Portfolio），并对多因子投资组合的绩效表现进行衡量，评价指标包括年化收益率、累计收益率、波动率、夏普比率、FF3 调整的 Alpha 收益、及 FF5 调整的 Alpha 收益。本文对根据机器学习模型预测收益率构造的多空组合的累计收益率进行作图，并对以上所有检验的显著性进行检验。

最后，本文将通过迪堡马里亚诺（Diebold-Mariano）检验统计量和计算机运行时间两个指标来综合考虑与比较多因子投资策略模型的预测能力以及运算速度。

1.5 本文的结构

本次毕业论文的行文框架为：

第一章为绪论部分，介绍本次论文的选题背景以及研究目的与意义、多因子量化投资与机器学习量化投资相关文献综述、研究内容和研究方法、并介绍本文的行文结构。

第二章介绍多因子投资策略方法与框架，通过投资策略模型指标介绍、多因子投资策略模型框架介绍和 Boosting 系列机器学习算法介绍三个部分来具体阐述量化投资、机器学习、多因子选股模型的相关理论和概念、模型评价指标（包括夏普比率（William,1964）[8] 、Fama Franch 三因子调整的 Alpha 收益和经过 Fama Franch 五因子调整的 Alpha 收益），并描述多因子模型的框架。

第三章为因子投资策略实证分析。首先对本次实证的数据集、数据选取与预处理进行详细的描述，包括对因子的分类及数量、数据滞后性处理、缺失值处理进行描述。其次对单因子投资策略模型进行分析，包括单因子描述性统计分析、单因子多空投资组合构造、因子收益率的计算等。再次，本章将使用多个 Boosting 系列的机器学习算法，如极限随机森林、梯度提升决策树、极端梯度提升树以及分布式高效提升器，来进行横截面超额收益预测模型的构建，并计算相关策略评价指标。本章将通过比较不同 Boosting 系列机器学习算法的样本外预测表现，统计整理各个模型的运行时间，对计算机运行性能进行评价。在稳健性检验的部分，本文构造市值加权投资组合的方式来检验模型算法的稳健性。

第四章为结论和展望，首先对本文的研究进行总结，并对目前研究的局限性以及基于本文的不足派生出来的未来可能的研究发展方向进行阐述。

2 多因子投资策略方法与框架

第二部分主要介绍本文所使用的衡量模型表现优劣的指标、多因子模型框架和 Boosting 系列机器学习算法。

2.1 投资策略模型评价指标介绍

2.1.1 夏普比率（Sharp Ratio）

夏普比率衡量的是单位风险承担水平上的风险溢价（William,1964）[8] 。它的内在含义是利险相随，承担更高的风险可以带来很大的收益，但是投资者仍然可以选择一定风险承担水平下，超额收益更高的资产，或者超额收益一定的情况下，带来的风险最小的资产。夏普比率是学界和业界最常用的投资策略模型绩效评价指标，它能综合代理系统性风险和非系统性风险，体现的是资产的总体风险。

夏普比率的公式如（2.1）所示：

$$\text{SharpRatio}=\frac{\sum_{i=1}^{n}(R_i-R_f)}{\sigma_{R_i}} \tag{2.1}$$

2.1.2 经 FF3/FF5 调整的 Alpha 收益

经过 Fama Franch 三因子调整的 Alpha 收益和经过 Fama Franch 五因子调整的 Alpha 收益是因子文献中常用的检验因子有效性的手段。Fama 和 French(1992)[2] 三因子模型包括市场风险溢价（Market Risk Premium）、市值因子（SMB）、和账面市值比（HML）这三个因子。在计算 FF3-alpha 时，将计算出来的预测的 t+1 期的股票的月度超额收益数据作为因变量，Fama Franch 三因子作为自变量进行回归。本文中 FF3 组合划分基于 FAMA2*3 组合划分方法；FF5 采用的是 FAMA2*2*2*2 投资组合划分方法。组合投资收益率的计算均采用总市值加权。回归式如公式（2.2）所示。

以上回归得到的截距项即经过 Fama Franch 三因子调整的 Alpha 收益，即 Fama Franch 三因子模型不能捕捉，但是可以被新的因子模型所捕捉的收益率部分。截距的 t 统计量即 FF3-Alpha 的显著性水平。回归的表达式如公式（2.2）所示：

$$r_p - r_\alpha = \alpha + \sum_{m=1}^{L} \lambda_m \beta_m \tag{2.2}$$

其中r_p为投资组合的收益、r_α为基准投资组合的收益、β_m为投资组合在因子 m

上的因子载荷、λ_m为因子 m 的风险溢价、L 为 3 或 4（取决于风险模型使用的是 FF3 还是 FF4）、α为经过风险调整的收益。

经过 Fama Franch 五因子调整的 Alpha 收益中五因子包括市场风险溢价因子（Market Risk Premium）、市值因子（SMB）、账面市值比（HML）、盈利能力因子（RMW）、投资模式因子（CMA）这五个因子，计算方式与经过 Fama Franch 三因子调整的 Alpha 收益类似。

2.1.3 迪堡马里亚诺统计量（Diebold Mariano test）

迪堡马里亚诺统计量用于衡量两个预测模型的准确性之间存在多大的差异。迪堡马里亚诺检验的原假设是两个预测模型之间不存在差异。与以往描述预测模型准确性差异的统计量不同，它放松了对预测误差的假设，预测误差可以满足非高斯分布、非零均值、序列相关等条件（Diebold, Mariano,1995）[36] 。

对于股票超额收益预测模型，时间上的序列相关性相对较弱，但是在横截面中的股票水平的序列相对更强，本文使用迪堡马里亚诺检验来对比两两模型之间的横截面水平上的平均预测误差是否合理。DM 检验统计量表达式如公式（2.3）所示：

$$d_{12,t+1} = \frac{1}{n3}\sum_{i=1}^{n3}((\hat{e}_{i,t+1}^{(1)})^2 - (\hat{e}_{i,t+1}^{(2)})^2) \tag{2.3}$$

其中，$(\hat{e}_{i,t+1}^{(1)})^2$和$(\hat{e}_{i,t+1}^{(2)})^2$代表两种方法中每种方法的预测误差，n3 代表的是测试集中的股票数目。那么，$\overline{d_{12}}$和$\widehat{\sigma_{\overline{d_{12}}}}$代表$d_{12,t}$在测试集上的均值和 Newey-West 标准差。这种调整的迪堡马里亚诺统计量是基于时间序列上相关性小但是存在个体差异的单个序列而创造出来的，它能满足偏正态分布的条件，给模型的两两比较提供一致的基准（Gu, Kelly, Xiu, 2018）[35] 。

2.2 多因子投资策略模型框架

基于面板数据截面回归的多因子回归模型的基本框架是将每个股票 t 期的因子数据作为回归模型的自变量，每个股票 t+1 期的月度超额收益数据（月度收益率减去月度无风险利率）作为回归模型的因变量。将回测期 T 期内所有的样本数据输入到模型中进行回归，得到预测的 t+1 期的每支股票的超额收益数据，根据这个预测结果划分多空组合在 t 期末进行投资，然后根据 t+1 期的实际月度收益率报数据计算投资组合的表现情况，包括投资组合的简单加权月度收益率、投资组合的

市值加权收益率、投资组合收益率在所有回测期的波动率，进而计算出投资组合的最大回撤（即投资组合净值从最高处回落到下一段时期内最低处的净值量占总净值增长的比率）和夏普比率（即单位风险承担上的风险溢价）。

其中，传统的线性多因子模型采用的是混合面板的方法，即将一个回测期内所有的股票数据单独作为样本输入到模型中，这就要求各个样本之间在截面上不存在显著差异，每个样本在时间序列水平上不存在显著差异。Fama Macbeth（1973）[7] 提出了更加稳健的衡量风险与收益关系的计量方法，即 Fama Macbeth 两步回归法。这种方法将 t 期每支股票的因子数据作为回归模型的自变量，将 t+1 期每支股票的月度收益率数据作为回归模型的因变量，在一个回测期内产生多个模型，将模型的回归系数简单加权平均作为最终预测模型的回归系数，而且回归系数更为稳健。在当时缺乏对面板数据，尤其是动态面板数据的研究的情况下，Fama Macbeth 的研究具有开创性，而且成为了后来股票市场多因子回归模型的基准模型。

除此之外，本文采用滚动窗口的方式进行样本外回测。相较于传统的划分方法（如留出法、交叉验证等）而言，滚动窗口划分方式在最大程度上保留了原始数据集合的时间序列特征，与现实中投资决策过程保持一致，也在很大程度上提高了数据的利用效率，保证市场信号具有一定的时效性。

将月度集成的截面回归模型与滚动时间窗口的回测方法相结合，本文模型具体构造方式即将一个滚动时间窗口内所有截面回归预测模型进行集成，对样本外月度数据进行预测，逐月滚动回测构造投资组合。具体示意图见下页图 2.1。

对于机器学习模型来说，与线性多因子模型相比，它有几个不同点：

首先，机器学习模型对输入模型的数据有更加严格的要求，即要求提前进行特征工程，包括数据预处理、缺失值填充，特征标准化等等。

其次，对于机器学习模型是否能直接处理面板数据缺乏系统性的理论研究作为支撑。目前的运用机器学习算法都是单纯分为截面有监督的回归或者是时间序列上有监督的回归；运用机器学习算法进行股票市场多因子模型构建的论文基本都采用混合面板的做法。在对样本进行相关检验之后，本文延用 Fama Macbeth 的思想进行机器学习多因子回归模型的构建，投资组合绩效表现良好。

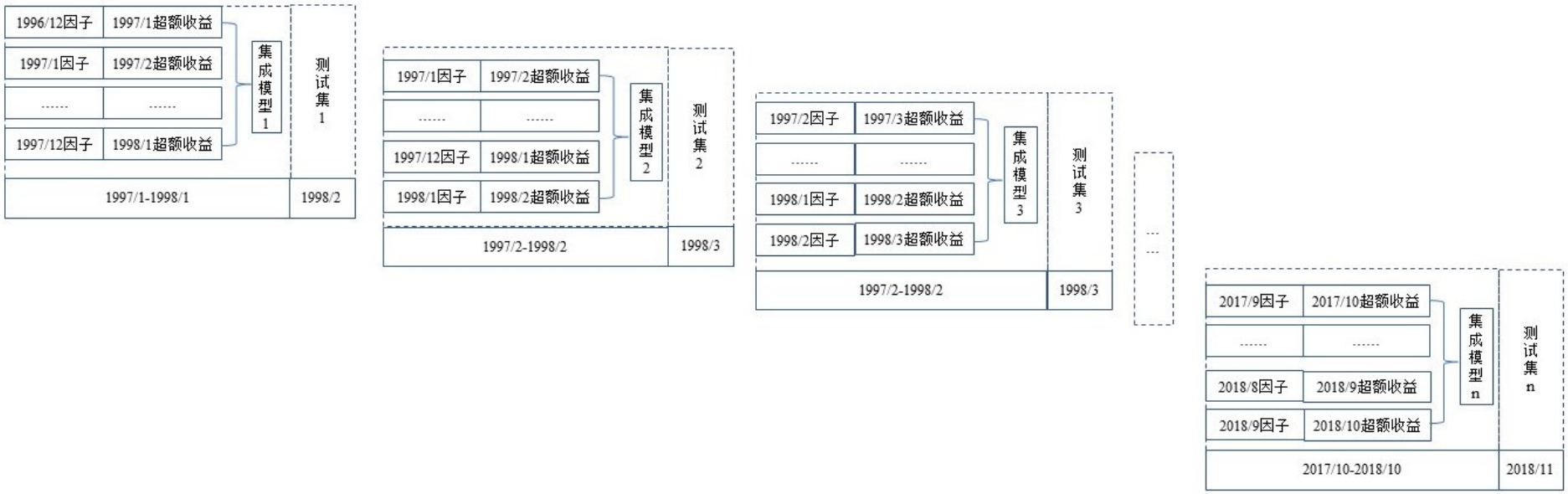


图 2.1 多因子投资策略模型框架

2.3 Boosting 系列机器学习算法介绍

2.3.1 极限随机森林（Extremely Randomized Forest）

决策树回归是针对连续型变量使用树结构来表示类划分，通过对特征集的识别、标记来得到空间的最优划分。与随机森林算法类似，极限随机森林（Extremely Randomized Forest）[9] 也是针对树回归算法的集成方法（Breiman, Leo, 2001）[40]（Breiman. etal. 1984）[41] ，这就意味着在分类器构造的过程中构造多个分类器会引入随机性。集成结果的预测是基于每个子分类器的平均结果。在极限随机森林算法中，在划分计算的过程中随机性被进一步加强。和随机森林算法一样，随机的特征子集会被使用，但是除了寻找最优阈值，对于每个特征而言最随机的阈值将被选择。这将导致模型的方差减小，偏差增大。

极限随机森林的主要超参数包括弱学习器的数量（n_estimators）、子分类器的最大特征数（max_features）和树的最大深度（max_depth）。一般而言，树的数量越多，模型越精确，但是运算的时间也更长。子分类器的最大特征数越大，方差越大，偏差越小。树的深度越大，模型越复杂。

本文所使用的极限随机森林的超参数为 Python3.7 的 sklearn0.19 版本中极限随机森林的默认参数，其中最大特征数为 10、子分类器的最大特征数为自动选择的参数、树的最大深度不受限制。

2.3.2 梯度提升决策树（Gradient Boosted Regression Trees）

梯度提升决策树（Gradient Boosted Regression Trees）是一种针对不同回归损失函数的 Boosting 集成学习算法（R. Meir and G. Rätsch, 2003）[42] 。Boosting 即依据错误率进行重抽样并将弱分类器集成为强分类器[10] [11] 。

在 GBRT 的迭代时，假设前一轮迭代得到的强学习器是$f_{t-1}(x)$,损失函数是$\mathrm{L}\left(y, f_{t-1}(x)\right)$。下一轮迭代的目标是得到一个决策回归树模型的弱学习器$h_t(x)$，使得本轮的损失$\mathrm{L}\left(y, f_t(x)\right)$=$\mathrm{L}\left(y, f_{t-1}(x)\right)$+$h_t(x)$最小。也即，本轮迭代得到的弱分类器使得样本的损失变得更小。

GBRT 回归算法的基本步骤有三大步：

（1） 初始化弱学习器 c

$$f_0(x) = \operatorname{argmin}(c) \tag{2.4}$$

（2） 对迭代轮数 t=1,2,…,T 有：

①对样本 i=1,2,…,m,计算负梯度

$$r_{ti} = [\frac{\partial L(y,f(x_i)}{\partial y,f_{t-1}(x)}]f(x) = f_{t-1}(x) \tag{2.5}$$

②利用((x_i,r_{ti})(i=1,2,…,m),拟合一棵 CART 回归树，得到第 t 棵回归树，其对应的叶子节点区域为$R_{tj}, j = 1,2,\ldots,J$。其中 J 为回归树 t 叶子节点的个数。

③对叶子区域 j=1,2,…J,计算最佳拟合值

$$\epsilon_{tj} = argmin(c)\sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x_i) + c) \tag{2.6}$$

④更新强学习器

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{I} c_{tj} I\left(x_i \in R_{tj}\right) \tag{2.7}$$

⑤得到强学习器 f(x)的表达式

$$f(x) = f_\tau(x) = \sum_{t-1}^{\tau}\sum_{j=1}^{J} c_{tj} I\left(x_i \in R_{tj}\right) \tag{2.8}$$

梯度提升决策树模型的主要参数分为两类，一类是梯度提升决策树模型框架参数，与具体的弱学习器无关，包括弱学习器的个数（n_estimators）、学习率（learning_rate，即每个弱学习器的权重缩减系数）、子采样率（subsample）；另一类是梯度提升决策树模型弱学习器参数，包括决策树划分时的最大特征数（max_features）和树的最大深度（max_depth）。

本文所使用的梯度提升决策树的超参数为 Python3.7 的 sklearn0.19 中梯度提升决策树的默认参数，其中最大特征数为 100、子分类器的最大特征数为自动选择的参数、树的最大深度为 3。

2.3.3 极端梯度提升树（XGBoost）

极端梯度提升树是梯度提升决策树方法的一种变形[12] 。其目标函数为：

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{i=1}^{n} \Omega(f_k) \tag{2.9}$$

其中 n 为样本的个数，(x_i)表示第 i 个样本，y_i和$\hat{y_i}$为第 i 个样本的真实值和预测值；K 为 CART 的个数，f_k表示第 k 个 CART，可看做从样本点到分数的映射；$L(y_i, \hat{y_i})$为损失函数，$\Omega(f_k)$为正则项。

在训练第 t 棵树时，相当于极小化第 t 棵树的目标函数：

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{i=1}^{n} \Omega(f_k)$$

$$= \sum_{i=1}^{n} L((y_i, \hat{y}_t^{(t-1)}) + f_k(x_i)) + \Omega(f_k) + const$$

$$= \sum_{i=1}^{n} L[(y_i, \hat{y}_t^{(t-1)}) + g_i f_t(x_i)) + \frac{1}{2} h_i {f_t}^2(x_i) + \Omega(f_k) + const \quad (2.10)$$

XGBoost的特别之处就是用损失函数的二阶泰勒展开来近似原来的损失函数，上述目标函数可近似为：

其中g_i和h_i分别为第 i 个样本点上损失函数 L 关于第二个变量的一阶和二阶偏导数。如果我们再知道f_t和Ω的表达式，就能得到第 t 棵树。

目标函数中正则项部分（相当于树的复杂度）如下定义：

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{J=1}^{T} \omega_j^2 \quad (2.11)$$

其中f_t表示第 t 棵树，T 表示该树的叶子节点的个数，ω_j表示第 j 个叶子节点上的分数；λ和γ为惩罚因子，越大表明对树的复杂度的惩罚力度越大。

对f_t细化，将树拆分成结构部分 q 和权重部分ω：

$$f_t(x) = \omega_{q(x)} \omega \in R^T, q: R^d \rightarrow \{1,2,\ldots,T\} \quad (2.12)$$

其中ω是一个 T 维向量，对应 T 个叶子节点上的分数；q 是一个映射，将样本点 x∈R^d映射到某一个叶子节点上。因此，只要确定了树的结构 q 和每个叶子节点上的得分ω，就能完全确定这棵树[12] 。

极端梯度提升树模型的框架参数与 GBRT 相同，不同的是它增加了两个弱学习器参数，构建树时的抽样比例（colsample_bytree）和决策树每层分裂时的列抽样比例（colsample_bylevel）。

本文所使用的极限梯度提升树的超参数为 Python3.7 的 xgboost 包中极限梯度提升决策树的自带参数，其中最大特征数为 100、树的最大深度为 3、构建树时的抽样比例为 1、决策树每层分裂时的列抽样比例为 1。

2.3.4 分布式高效提升器（lightBGM）

分布式高效提升器是微软 2016 年开源的 Boosting 算法，与极端梯度提升树相比计算速度更快[14] 。它将梯度提升决策树与基于梯度的单项抽样方法（Gradient-based One-Side Sampling，即 GOSS）和互斥特征捆绑方法（Exclusive Feature Bundling，即 EFB）结合起来（Ridgeway, 2007）[13] （Qi Meng, 2016）[43] 。有了 GOSS，模型将会派出梯度小的大部分数据，只使用剩下的数据来估计信息增益。有了 EFB，模型可以自动捆绑互斥的特征来减少特征的数量。相比传统的梯

度提升决策树算法，在相同的准确度下，分布式高效提升器算法可以提升计算准确率约 20 倍（Qi Meng, etal, 2016）[43] （Huan Zhang, 2018）[44] 。

3 因子投资策略实证分析

3.1 数据与处理

本文实证研究中输入模型的因子数据的时间区间为1996年12月至2018年10月。数据采取这个时间区间的主要原因是自1996年12月16日开始，深交所和上交所对上市股票实行涨跌停板限制，即最大日涨跌幅度为 10%。后续研究表明，这一交易机制对股票市场研究存在一定的影响（吴林祥等，2013）[45] ，所以本文的起始时间为1996年12月底。

本文的股票池为全部A股，并对股票池进行了细致的处理，如剔除金融股（按照证监会2012年行业分类标准），ST以及ST*股，所有股票停牌期间数据，以及所有股票上市一年内的数据。这样处理的主要原因包括：金融行业与实体行业公司基本面指标差异较大，ST以及ST*股存在退市风险会对研究结果产生影响，股票停牌期间容易产生股票价格异常波动，股票上市一年内存在 IPO 溢价效应（金超,柯昌隆，2011）[46] 。

数据频率为股票月度数据，对于数据频率比月度频率更加稀疏的数据，如季度公布的财务数据指标，为了避免使用未来数据，使用最近会计期间内可得的数据进行填充，对截面上所有公司而言该数据填充方式相同。比如，t-1年的年报数据最迟公布时间为t年4月底，半年报公布时间最迟为t年8月底，三季报公布时间最迟为t年10月底。本文用t-1年年报数据填充t年5月至8月的月度数据，t年的半年报数据填充t年9月至10月的月度数据，t年三季报数据填充t年11月至t+1年4月的月度数据，以此类推。

为了使模型更好地拟合，本文采取了通过Copula函数在截面将所有的因子进行标准正态化的方法进行数据预处理。Copula 函数考虑了多元自变量之间的关系对它们统一映射到多元标准正态分布函数中，这种方法被广泛运用于投资组合管理（Low. etal,2013）[47] 。

从截面上看，本文所采用的数据一共 262 个截面，每个截面上的解释变量为96 个，是一个不均衡的面板。考虑到机器学习模型对特征的特殊要求，预测模型的拟合要求对截面上的因子缺失值进行填充。考虑到本文综合使用了因子以及行业均值调整后的因子（如行业调整账面市值比（BM_ia）、行业均值调整后的现金流价格比（CFP_ia）、行业调整后资本支出变动百分比（pchcapx_ia）、行业均值调

整后的雇员人数变动（chempia））、本文采用的缺失值填充方式是全部填充为零，起始时间晚于 1996 年 12 月的因子在产生之前的缺失值全部填充为零。

本文日交易数据来源于 Wind 数据库（万德数据库），财务报表数据来源于 CSMAR 数据库（国泰安数据库）。

3.2 单因子投资策略分析

本文采用的中国 A 股市场量化异象因子数据库子中共有 100 个因子，根据单因子文献，将它们分为交易摩擦类因子（共 21 个）、动量因子（共 6 个）、价值因子（共 10 个）、成长因子（共 29 个）、盈利性因子（共 18 个）、财务流动性因子（共 10 个）、其他类因子（共 6 个）。

生效时间晚于 1996 年 12 月的因子中，有两个开始于 1997 年 6 月，有 3 个开始于 1998 年 9 月，有 5 个开始与 1998 年 10 月，有 1 个开始于 2000 年 4 月，有 1 个开始与 2000 年 5 月，有 3 个开始于 2000 年 6 月，有 3 个开始于 2002 年 2 月，根据本文的数据预处理方式，生效前的数据全部填充为 0。

剔除收入增加期数（nincr）、发放股利指标（divi）、停止股利指标（divo）、有罪的股票（sin，代理烟草和酒精行业）这四个无法进行单因子检验的虚拟变量因子，以及首次公开发行因子（IPO 因子）之后，本文对单因子进行检验，计算它们的因子收益率、波动率、夏普比率、最大回撤、累计收益率这五项指标。

其中，因子收益率的计算即使用数据处理部分得到的因子月度数据，将对应可交易股票的因子按从高到低进行排序，平均分为十组，取第一组和第十组的数据分别按照等权重、市值加权两种方式构造相应的投资组合，采用 10-1 多空组合的方式得到月度因子收益率数据。

结果表明，年化收益率的 t 统计量的绝对值大于 3 的因子（Green. etal，2013）[19] 中，简单加权投资组合共选出 12 个因子，其中交易摩擦类因子共 11 个，市值加权投资组合共选出 5 个因子，全部为交易摩擦类因子；两种组合方式中表现最优的因子均为市值因子（size），其年化收益率为 21.26%。从中国股票市场的未来发展来看，市值因子的表现过于优异会对单因子的进一步研究产生限制作用。随着各行业的兴起，市场竞争的加剧，以及市场竞争格局的逐步稳定，小市值公司抢占市场竞争优势的可能性越来越小，未来不一定会重演历史的小市值效应。

从简单加权多空组合上看，年化收益率排名前十的因子中，交易摩擦类因子

有 7 个，成长因子有 1 个、动量因子有 1 个、盈利因子有 1 个。从市值加权多空组合上看，年化收益率排名前十的因子中，交易摩擦类因子有 5 个，成长因子有 2 个、动量因子有 1 个、盈利因子有 2 个。

在简单加权平均构造的投资组合的单因子的显著性水平上，有 30 个因子在 95%的置信区间上显著，其中有 13 个交易摩擦类因子，占该类别的 61.90%；3 个动量因子，占该类别的 50%；4 个价值类因子，占该类别的 44.44%；3 个财务流动性因子,占该类别的 33.33%；5 个成长因子，占该类别的 17.86%。在市值加权平均构造的投资组合的单因子的显著性水平上，有 26 个因子在 95%的置信区间上显著，其中有 11 个交易摩擦类因子，占该类别的 52.38%；有 3 个动量因子，占该类别的 50.00%;有 6 个盈利因子,占该类别的 26.09%;有 5 个成长因子,占该类别的 17.86%;有 1 个财务流动性因子，占该类别的 11.11%。

无论是从年化收益率排名还是从因子显著性上来看，单因子筛选因子的结果表明单因子模型对于交易摩擦类因子更为敏感，在传统的单因子模型的结论下倾向于认为中国 A 股市场仍然是由噪声交易者主导的市场。

表 3.1 简单加权投资组合单因子结果分析[2]

因子	年化收益率	T 统计量	波动率	最大回撤	累计收益率	夏普比率	因子类型
size	-0.2126	-3.8701	0.2567	0.4336	50.2997	0.7218	交易摩擦类因子
std_dvol	-0.1818	-8.7984	0.0966	0.1085	47.2649	1.5999	交易摩擦类因子
lagretn	-0.1517	-3.8841	0.1824	0.3678	19.0395	0.6814	交易摩擦类因子
rd_mve	0.1393	5.4255	0.1200	0.1853	17.8393	0.9333	成长因子
volumed	-0.1165	-5.9414	0.0916	0.0889	11.5936	0.9731	交易摩擦类因子
std_turn	-0.1162	-6.5113	0.0834	0.1260	11.6868	1.0656	交易摩擦类因子
illq	-0.0921	-4.2123	0.1022	0.1836	6.6755	0.6339	交易摩擦类因子
chfeps	0.0844	4.7377	3.9399	0.0729	0.1669	0.8270	成长因子
aeavol	-0.0667	-3.6105	0.0863	0.8078	3.9638	0.4559	交易摩擦类因子
retnmax	-0.0599	-4.5582	0.0614	0.0940	3.5558	0.5304	交易摩擦类因子
idvol	-0.0474	-3.2926	0.0673	0.0798	2.6841	0.2982	交易摩擦类因子

[2] 按照显著性水平排名，且 t 的绝对值大于 3。

表 3.2 市值加权投资组合单因子结果分析

因子	年化收益率	t-statistics	波动率	最大回撤	累计收益率	夏普比率	因子类型
std_dvol	-0.1318	-5.5332	0.1113	0.1676	15.4655	0.9385	交易摩擦类因子
std_turn	-0.1026	-4.5128	0.1062	0.1707	8.2969	0.7084	交易摩擦类因子
volumed	-0.0979	-4.3786	0.1044	0.1737	7.5293	0.6753	交易摩擦类因子
size	-0.2167	-3.6265	0.2792	0.4999	48.1711	0.6782	交易摩擦类因子
aeavol	-0.0610	-3.5684	0.0798	0.2031	3.5351	0.4211	交易摩擦类因子

3.3 多因子投资策略分析

本文的多因子框架中最终采取了中国 A 股市场量化异象因子数据库子中的 96 个因子，剔除了预期每股收益的变化（chfes）、未预期收益（sue）、收益预测（sfe）、分析师人数变化（chnanalyst）、涉及股票的分析师人数（nanalyst）这五个因子，因为它们起始的时间过晚，缺失值过多，如果将这几个因子纳入，本文所采取的缺失值填充方法会对结果造成较大影响。剔除新股发行（IPO）因子，该因子为公司上市当年设为 1，其余截面内设为 0；因为剔除了上市第一年的数据，所以所有的可利用的数据为 0，在本文的分析框架下无法被纳入多因子模型框架中去。所以，本文最终的多因子模型只选取了 96 个因子。

该模型对面板数据进行横截面的回归预测，即将 3571 支股票 1 至 t 月的 n 个因子值作为模型的自变量，将 1 至 t 月的每支股票的收益率作为模型的因变量，调仓频率为月，模型开始训练时，1 至 t 月作为模型的训练集和调参集。采用滚动窗口的方式，逐月向前滚动训练、调参，滚动窗口的长度采用 12 个月、36 个月、60 个月。将线性回归模型作为基准模型进行对比，采用一系列的 Boosting 模型进行滚回预测，具体介绍见第二章第二节多因子投资策略模型框架。

本文计算根据机器学习模型在测试集上预测的每支股票的收益率作为机器学习模型输出的因子值，对该值进行排序逐月滚动构造投资组合，包括多空投资组合、多投资组合和空投资组合，并对全部 A 股股票池计算所构造因子组合的年化收益率、累计收益率、波动率、夏普比率、FF3 调整的 Alpha 收益、及 FF5 调整的 Alpha 收益，对根据机器学习模型预测收益率构造的多空组合的累计收益率进行作图，并对其显著性进行检验。

针对不同的 Boosting 算法机器学习模型，本文通过计算样本外迪堡马里亚诺检验统计量对模型的预测精确度进行两两比较[36]；当比较多因子投资策略模型的运算速度和运算效率时，本文用计算机运行时间、夏普比比上运行时间这两个指

标来考虑。能够构造出夏普比率高的投资策略的模型可能耗时长，计算速度慢，在实盘交易中，等策略运行出来市场上的交易机会可能已经消失；而耗时短、计算速度快的算法可能投资策略的绩效不尽人意；综合考虑来说，运行速度相对快，夏普比率相对高，夏普比比上运行时间高的模型对应的算法是最理想的算法。

3.3.1 混合面板模型 F 检验

首先，根据 F 检验的结果，来判断本文所使用的面板数据是否适用于混合面板模型。

检验的原假设为对于不同横截面模型截距项相同（即可以建立混合估计模型）。计算的各截面个体固定效用模型的残差平方和为 0.00437，记为 SSEr，混合截面的残差平方和为 0.00845，记为 SSEu，F 统计量的表达式为

$$F=\frac{(SSEr-SSEu)/(T+k-2)}{SSEu/(NT-T-k)} \tag{3.1}$$

其中模型中含有 k 个解释变量、T 个时期、每个截面上 N 个样本，计算可得 F 统计量为 22200，拒绝可以使用混合面板模型的原假设，简单地使用混合面板模型不存在合理性，所以本文使用 Fama Macbeth（1973）[7] 的回归构造方法，并将该方法迁移到机器学习预测模型上。

该检验验证了本文模型构造方式的合理性，具体的模型构造方式可以参见第二章第二节多因子投资策略模型框架。

3.3.2 多因子投资策略实证结果分析

本文根据机器学习算法预测的月度超额收益逐月构造 10-1 多空投资组合、10 多头投资组合、1 空头投资组合，通过计算投资组合的月度超额收益率、月度收益率、年化收益率、年化收益率的波动率、1998 年 1 月至 2018 年 10 月期间平均的年化收益率、投资组合的夏普比率、投资组合经过 FF3 调整后的 Alpha 收益、以及投资组合经过 FF5 调整后的 Alpha 收益来对各个算法的效果进行投资学上的对比分析。

结果表明，多因子模型 12 月多空组合相比平均的最优的前 10 个单因子模型年化收益率能提高 29.91%、夏普比率相对提高 0.49；机器学习 Boosting 集成算法多空组合相比 Fama Macbeth 回归年化收益率能提高 10.04%、夏普比率相对提高 0.55，其预测模型的公式如公式（3.2）所示。

$$R_{ensamble,t}=\frac{1}{4}\sum_{i=1}^{4}R_{i,t} \tag{3.2}$$

更高级的 Boosting 算法模型如极端梯度提升树和分布式高效提升器的结果普遍优于较简单的 Boosting 算法模型如极限随机森林，比如 60 个月的极端梯度提升

树和分布式高效提升器模型的年化收益率分别为 62.58%和 64.63%，60 个月的极限随机森林模型的年化收益率为 50.49%。

从模型的显著性水平上看，多空组合和多头组合的所有的指标都显著，空头组合的大部分指标都显著，且多空组合的指标显著性普遍高于多头组合，符合本文对于投资组合绩效表现的预期。

从市场信号的角度分析出发，时间窗口越长，模型所能接受的信号越多，实证结果表明时间窗口越长，各指标不显著的数目越少，所有投资组合的 FF3 和 FF5 均显著，所有多空组合和多头组合的年化收益率、夏普比率均显著；除了 Fama Macbeth 回归和梯度提升决策树在 36 个月的表现最好以外，其余模型均呈现时间窗口越长表现越优，具体的对比结果可参照图 3.1 FM 与 Boosting 算法多空组合对比情况。结果最好的模型为 60 个月的分布式高效提升器模型，其构造的简单加权的多空投资组合夏普比率可达到 2.52、年化收益率可达到 64.63%，FF3-Alpha 收益可以达到 69.15%，FF5-Alpha 收益 63.80%，表现优于沪深 300 指数在相同时间窗口内的表现。

本文还计算了多空投资组合的回测期内的净值图，表现最好的 60 个月的分布式高效提升树模型的净值比例可以达到 3874 倍。净值表示的是期初投资一单位货币，在回测期内维持投资观念或者投资策略不变并且将所有收益用于再投资，投资期结束时的投资组合净值，也即投资组合的累积收益率。从时间窗口上看，时间窗口越长，各模型在投资组合净值上的表现差异越大，图 3.1 展示了投资组合净值去对数后的比例变化图。

表 3.3 投资策略绩效分析（简单加权平均）

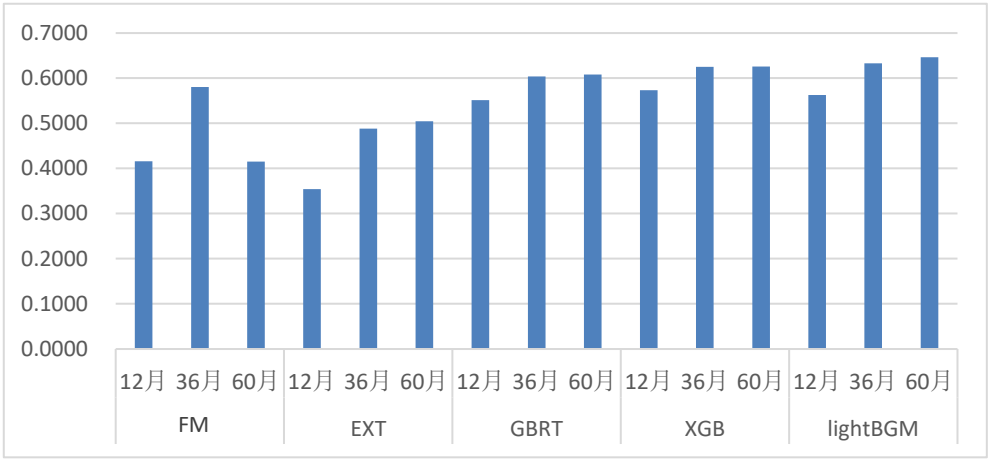
简单加权平均	12 个月					36 个月					60 个月				
模型	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM
多空组合年化收益率	0.4161	0.4289	0.5512	0.5733	0.5626	0.5807	0.4883	0.6040	0.6252	0.6332	0.4150	0.5049	0.6080	0.6258	0.6463
多空组合夏普比率	1.3257	1.3401	1.9944	2.1504	1.9669	2.1890	1.9414	2.3050	2.3662	2.4446	1.2388	1.9571	2.2839	2.3707	2.5248
多空组合 FF3-alpha	0.4291	0.4437	0.5686	0.5894	0.5817	0.6210	0.4922	0.6509	0.6699	0.6809	0.4152	0.5428	0.6556	0.6724	0.6915
多空组合 FF5-alpha	0.4378	0.4083	0.5222	0.5430	0.5400	0.5497	0.4940	0.5807	0.5915	0.6075	0.4307	0.5036	0.6059	0.6189	0.6380
多头组合年化收益率	0.2617	0.2562	0.3063	0.3162	0.3213	0.3359	0.2450	0.3299	0.3464	0.3517	0.2526	0.2841	0.3307	0.3453	0.3571
多头组合夏普比率	0.1490	0.1340	0.2854	0.3132	0.3252	0.3955	0.2584	0.3758	0.4168	0.4331	0.1479	0.2309	0.3639	0.4025	0.4301
多头组合 FF3-alpha	0.2715	0.1298	0.1852	0.1948	0.2012	0.2342	0.2589	0.2313	0.2451	0.2487	0.2496	0.1844	0.2413	0.2546	0.2642
多头组合 FF5-alpha	0.3064	0.1177	0.1702	0.1791	0.1847	0.2090	0.2738	0.2034	0.2134	0.2150	0.3196	0.1729	0.2260	0.2373	0.2432
空头组合年化收益率	0.0570	0.0387	-0.0336	-0.0457	-0.0300	-0.0452	-0.0438	0.4555	-0.0793	-0.0821	0.0387	-0.0196	-0.0761	-0.0794	-0.0881
空头组合夏普比率	-0.4512	-0.4946	-0.7137	-0.7550	-0.6999	-0.7337	-0.6069	-0.8018	-0.8162	-0.8100	-0.4675	-0.6186	-0.7769	-0.7882	-0.8076
空头组合 FF3-alpha	0.0525	-0.1038	-0.1733	-0.1845	-0.1704	-0.1870	-0.0335	-0.2198	-0.2251	-0.2324	0.0359	-0.1569	-0.2129	-0.2163	-0.2258
空头组合 FF5-alpha	0.0791	-0.0801	-0.1415	-0.1534	-0.1449	-0.1412	-0.0208	-0.1779	-0.1787	-0.1931	0.0897	-0.1300	-0.1792	-0.1808	-0.1940

表 3.4 投资组合绩效指标的显著性分析（简单加权平均）

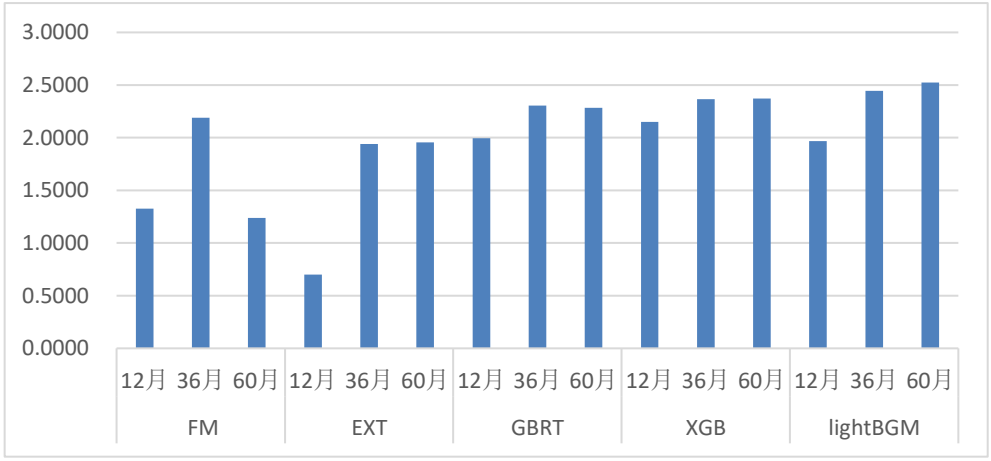
简单加权平均	12 个月					36 个月					60 个月				
模型	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM
多空组合年化收益率	42.6843	41.8634	42.6843	42.6843	49.9143	50.0193	47.6612	51.6244	52.1245	53.5350	34.0820	46.1208	48.3880	49.5287	51.9676
多空组合 FF3-alpha	12.3185	12.1252	14.7813	15.4887	14.4453	16.7390	13.4606	17.3418	17.4733	17.9682	9.6232	15.9039	16.9819	17.2506	17.9761
多空组合 FF5-alpha	12.1401	11.0508	13.4858	14.2128	13.2416	14.7284	12.7803	15.0933	15.3033	15.9382	9.4046	14.0689	15.0010	15.4346	16.1136
多头组合年化收益率	12.2668	12.1263	12.2668	12.2668	15.0615	14.6048	10.4969	14.2580	14.7428	15.0120	10.2934	11.2055	13.1687	13.6667	13.9592
多头组合 FF3-alpha	3.5917	5.0513	7.1656	7.5024	7.1029	9.2218	3.1114	8.5324	8.7840	8.8756	2.9100	7.8141	9.2531	9.5623	9.6256
多头组合 FF5-alpha	3.9184	4.2034	6.0131	6.2985	6.0292	7.5467	3.1068	6.8923	7.0922	7.1427	3.5349	6.5728	7.8624	8.1715	8.1431
空头组合年化收益率	2.6380	1.7564	2.6380	2.6380	-1.3771	-2.0347	-1.9161	-3.2742	-3.4836	-3.5414	1.5802	-0.7786	-3.0242	-3.1614	-3.4866
空头组合 FF3-alpha	0.6818	-4.6040	-7.4085	-8.0152	-7.6282	-8.3142	-0.4108	-9.4224	-9.5779	-10.3555	0.4152	-6.7418	-8.3339	-8.5176	-9.2987
空头组合 FF5-alpha	0.9910	-3.1458	-5.4912	-6.0433	-5.7957	-5.3280	-0.2416	-6.4741	-6.5411	-7.2182	0.9829	-4.4606	-5.8043	-5.9024	-6.6018

表 3.5 Boosting 集成算法策略结果分析（简单加权平均）

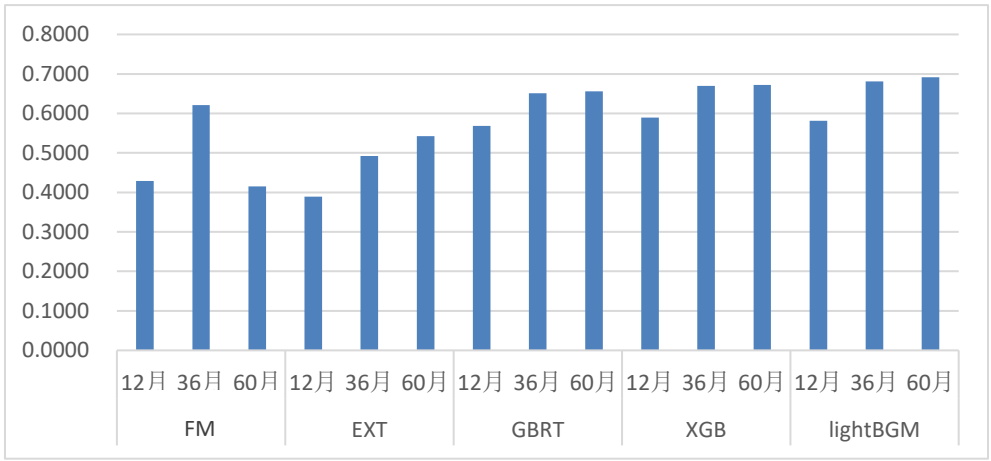
简单加权平均	12 个月			36 个月			60 个月			各时间窗口平均		
模型	FM	Boosting	变化情况	FM	Boosting	变化情况	FM	Boosting	变化情况	FM	Boosting	变化情况
多空组合年化收益率	0.4161	0.5290	0.1129	0.5807	0.5877	0.0070	0.4150	0.5962	0.1813	0.4706	0.5710	0.1004
多空组合夏普比率	1.3257	1.8630	0.5373	2.1890	2.2643	0.0753	1.2388	2.2841	1.0453	1.5845	2.1371	0.5526
多空组合 FF3-alpha	0.4291	0.5459	0.1167	0.6210	0.6235	0.0025	0.4152	0.6406	0.2254	0.4884	0.6033	0.1149
多空组合 FF5-alpha	0.4378	0.5034	0.0655	0.5497	0.5684	0.0188	0.4307	0.5916	0.1609	0.4727	0.5545	0.0817
多头组合年化收益率	0.2617	0.3000	0.0383	0.3359	0.3182	-0.0177	0.2526	0.3293	0.0768	0.2834	0.3159	0.0324
多头组合夏普比率	0.1490	0.2645	0.1155	0.3955	0.3710	-0.0245	0.1479	0.3568	0.2090	0.2308	0.3308	0.1000
多头组合 FF3-alpha	0.2715	0.1778	-0.0938	0.2342	0.2460	0.0118	0.2496	0.2361	-0.0135	0.2518	0.2200	-0.0318
多头组合 FF5-alpha	0.3064	0.1629	-0.1435	0.2090	0.2264	0.0174	0.3196	0.2198	-0.0998	0.2784	0.2030	-0.0753
空头组合年化收益率	0.0570	-0.0176	-0.0746	-0.0452	0.0626	0.1078	0.0387	-0.0658	-0.1045	0.0168	-0.0069	-0.0238
空头组合夏普比率	-0.4512	-0.6658	-0.2146	-0.7337	-0.7587	-0.0250	-0.4675	-0.7478	-0.2803	-0.5508	-0.7241	-0.1733
空头组合 FF3-alpha	0.0525	-0.1580	-0.2105	-0.1870	-0.1777	0.0093	0.0359	-0.2030	-0.2389	-0.0329	-0.1796	-0.1467
空头组合 FF5-alpha	0.0791	-0.1300	-0.2090	-0.1412	-0.1426	-0.0014	0.0897	-0.1710	-0.2607	0.0092	-0.1479	-0.1571



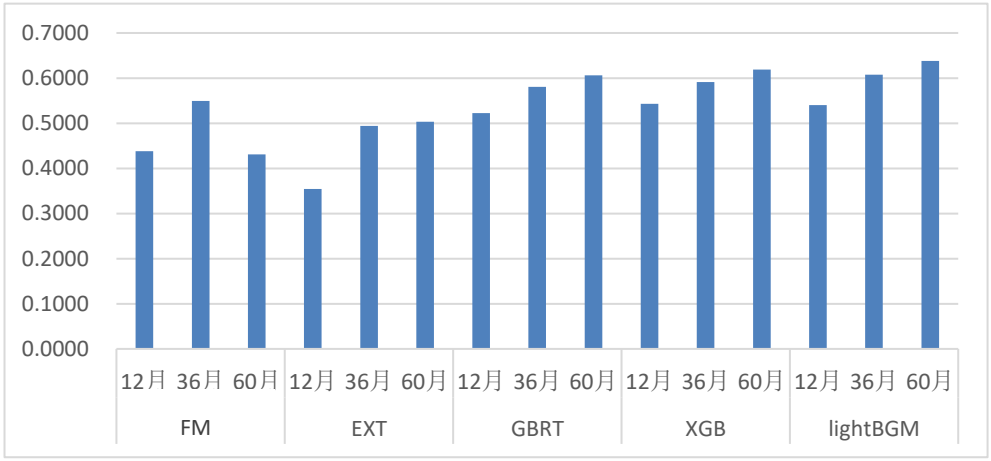
(a)多空组合年化收益



(b)多空组合夏普比率

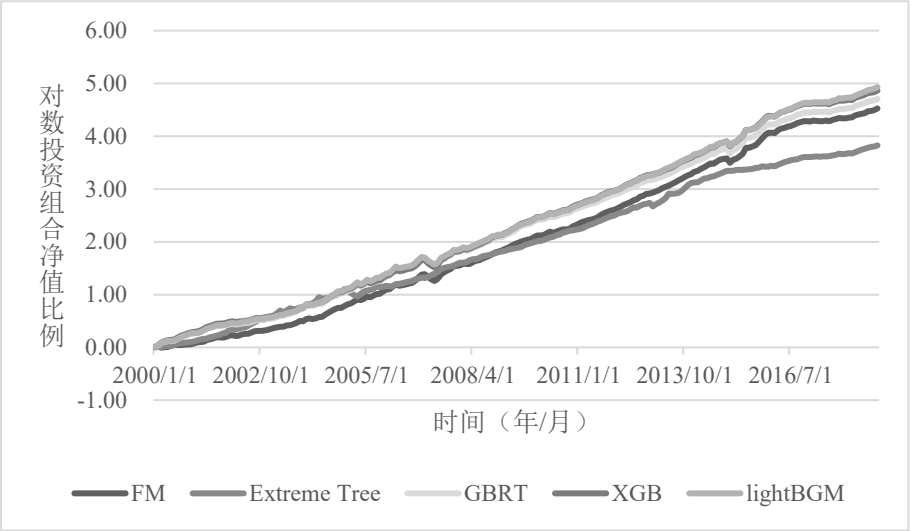


(c)多空组合 FF3-Alpha 收益

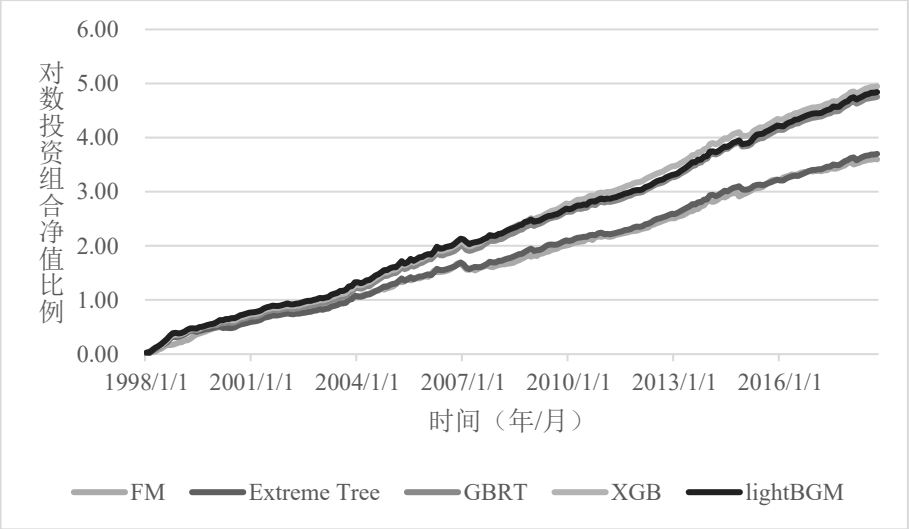


(d)多空组合 FF5-Alpha 收益

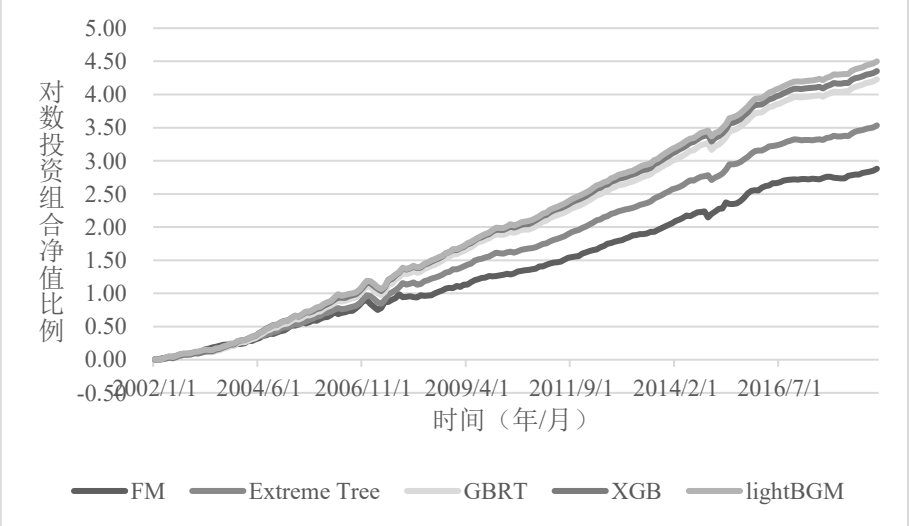
图 3.1 Fama Macbeth 与 Boosting 算法多空组合对比情况



(a)时间窗口为 12 个月



(b)时间窗口为 36 个月



(c)时间窗口为 60 个月

图 3.2 投资组合对数净值比例图（简单加权）

3.3.3 投资策略下的因子重要性

本文还计算了模型拟合之后得到的因子重要性程度，线性回归用回归斜率来代理，Boosting 系列算法用特征重要性程度来代理。Friedman（2001）[10] 提出，特征的全局重要性由该特征在单棵树中的重要性的平均值来衡量。本文通过计算各个模型在各个截面上的特征重要性，平均得到算法总体的特征重要性，并通过Z-值标准化（每个数据减去该列数据的均值再除以该列数据的标准差）的方法使得各个模型的特征重要性可以进行相对的比较。同时，由于 Boosting 算法在处理连续性变量和虚拟变量时存在差异，在考虑特征重要性的比较时将会剔除收入增加期数（nincr）、发放股利指标（divi）、停止股利指标（divo）、有罪的股票（sin，代理烟草和酒精行业）这几个虚拟变量因子。

在比较各类别的因子重要性程度时，对每一类各因子的处理后的因子重要性指标进行加权平均。其他类因子不是虚拟变量的只有一个，不具有参考意义，不参与比较。表 3.6 中单元格内数值表示该类指标处理后的因子重要性指标平均值，深浅程度表示在总体中大小的排名。

研究结果表明，极限树模型对因子的类别最不敏感，其次是 36 个月窗口的梯度提升决策树；Fama Macbeth 和分布式高效提升树对交易摩擦类因子和动量因子最为敏感；极端梯度提升树对财务流动性因子最为敏感；12 个月窗口的梯度提升决策树和 60 个月窗口的梯度提升决策树对交易摩擦类因子、动量因子、财务流动性因子较为敏感。

从时间窗口上分析，除了 36 个月窗口的梯度提升决策树与 12 个月、60 个月窗口的同类模型存在较大差异，其他算法在各个时间窗口上在因子重要性程度上的表现没有明显差异。

相比于单因子模型的表现排名，多因子模型能更好地识别中国股票市场上交易摩擦类因子以外的其他类因子，比如动量因子、财务流动性因子、成长类因子和盈利类因子。Fama Macbeth 挑选因子的结果和单因子模型比较类似，对于交易摩擦类因子赋予了较大的权重。总体来说，Boosting 系列模型对交易摩擦类因子、动量因子、财务流动性因子这三类因子最为敏感，相比为 Fama Macbeth 回归，Boosting 模型能更好更均衡地识别因子有效性，梯度提上决策树和极端梯度提升树能更好地识别财务流动性因子，分布式高效提升器能更好地识别动量因子。

从单个因子的结果来看，本文统计了单个因子层面各时间窗口各多因子模型共 15 个模型因子重要性情况，将单个模型中因子重要性程度排名前 20 的因子计数，15 个模型的情况进行累计后再次排名得到表 3.7。其中，收益公告异常交易量（aeavol）、12 个月动量（mom12）、公司年龄（age）、销售增长（Sgr）、投入资本回报（roic）被超过 80%的模型所选中，它们分别属于交易摩擦类因子、动量因子、交易摩擦类因子、成长因子以及盈利因子。这与单纯的单因子选择的结果差异较大，说明多因子模型可以更好地识别因子之间的相关关系，即在多个因子的共同作用下，模型可以识别出在单因子策略下表现欠佳的因子，捕捉多个变量之间复杂的非线性关系。

表 3.6 各算法的因子重要性程度

模型[3]	交易摩擦类因子	动量因子	价值因子	成长因子	盈利因子	财务流动性因子
FM-12	1.1310	0.6596	0.4202	0.3848	0.4706	0.3566
FM-36	0.8777	0.7195	0.3448	0.4212	0.6953	0.2760
FM-60	0.9628	0.8199	0.3299	0.4107	0.5432	0.3491
EXT-12	0.2052	0.2542	0.2091	0.2053	0.2182	0.2210
EXT-36	0.2172	0.2586	0.2087	0.2134	0.2294	0.2235
EXT-60	0.2251	0.2582	0.2181	0.2170	0.2264	0.2310
GBRT-12	0.6794	0.8131	0.4318	0.5353	0.4969	0.9421
GBRT-36	0.2172	0.2586	0.2087	0.2134	0.2294	0.2235
GBRT-60	0.7070	0.7804	0.4262	0.5390	0.5125	0.8815
XGB-12	0.3800	0.2705	0.2828	0.4711	0.3529	0.9072
XGB-36	0.3987	0.2844	0.2806	0.4716	0.3697	0.8919
XGB-60	0.4145	0.2827	0.2908	0.4908	0.3817	0.8982
lightBGM-12	0.9032	1.4929	0.3670	0.4658	0.5734	0.3892
lightBGM-36	0.9019	1.4272	0.3753	0.4794	0.5853	0.4174
lightBGM-60	0.8853	1.4216	0.3901	0.4865	0.5785	0.4305

表 3.7 模型因子筛选结果

因子排名	因子简称	被选中的次数	因子中文名称	因子类型
1	aeavol	15	收益公告异常交易量	交易摩擦类因子
2	mom12	15	12 个月动量	动量因子
3	age	12	公司年龄	交易摩擦类因子
4	Sgr	12	销售增长	成长因子
5	roic	12	投入资本回报	盈利因子
6	mom6	11	6 个月动量	动量因子
7	lagretn	11	短期反转	动量因子
8	betasq	10	系统性风险的平方	交易摩擦类因子
9	momchg	9	动量变化	动量因子
10	SgINVg	9	营业收入与存货增长率的差	成长因子
11	herf	9	行业销售集中度	其他
12	vol	8	总波动率	交易摩擦类因子
13	nincr	8	收入增加期数	盈利因子
14	CFdebt	8	现金流负债比	财务流动性因子
15	beta	7	系统性风险	交易摩擦类因子
16	std_dvol	6	交易额的波动率	交易摩擦类因子
17	SG	6	营业收入增长率	成长因子
18	egr	6	股东权益变化	成长因子
19	grCAPX	6	资本支出变化	成长因子
20	pchsale_pchinvt	6	销售增长减存货增长	成长因子

[3] 模型为算法加上对应的滚动时间窗口的长度组成。例如，12 月窗口的 Fama Macbeth 回归对应的缩写为 FM-12。

3.3.4 多因子投资模型预测能力比较

本文的多因子预测模型的对比通过计算迪堡马里亚诺检验统计量来进行。迪堡马里亚诺检验统计量的结果表明，从时间上看，Fama Macbeth 回归与极限随机森林在 137 个月份上不存在显著的差异；Fama Macbeth 回归与梯度提升决策树在 137 个月份上不存在显著的差异；Fama Macbeth 回归与极限梯度提升树在 138 个月份上不存在显著的差异；Fama Macbeth 回归与分布式高效提升器在 138 个月份上不存在显著的差异；极限随机森林与梯度提升决策树在 243 个月上不存在显著差异；极限随机森林与极端梯度提升树在 244 个月上不存在显著差异；极限随机森林与分布式高效提升器在 241 个月份上不存在显著的差异；梯度提升决策树与极端梯度提升树在 245 个月上不存在显著差异；极端梯度提升树与分布式高效提升器在 245 个月上不存在显著差异；梯度提升决策树与分布式高效提升器在 247 个月上不存在显著差异。

从迪堡马里亚诺检验统计量的假设检验结果上看，极限随机森林、梯度提升决策树、极端梯度提升树、分布式高效提升器这四个模型在预测效力上基本不存在差异，但是与简单线性回归的结果均存在差距。这与美国市场结果（Gu. etal，2018）[35] 相比相差较大。他们全面验证了各种类型的机器学习算法在资产定价领域的表现，且实证结果表明，简单线性方法和机器学习算法存在显著差异，线性机器学习算法如岭回归、LASSO 回归与非线性机器学习算法存在显著差异。

表 3.8 模型的预测能力比较：迪堡马里亚诺检验[4]

	FM	ExtrameTrees	GBRT	XGB	lightBGM
FM	0.0000	0.5020	0.4978	0.4919	-0.5066
	1.0000	**0.6764**	**0.6793**	**0.6822**	**0.6789**
ExtrameTrees		0.0000	0.0218	-0.0105	-0.0263
		1.0000	0.9795	0.9807	0.9764
GBRT			0.0000	-0.0097	-0.0057
			1.0000	0.9846	0.9912
XGB				0.0000	-0.0193
				1.0000	0.9840
lightBGM					0.0000
					1.0000

[4] 该表格每一栏第一行为迪堡马里亚诺检验（Diebold-Mariano）统计量的值，第二栏假设检验的 P 值。该检验的原假设为两个模型的预测结果不存在显著差异，加粗的部分为两个模型的预测结果在 40%以上的水平上存在差异。

3.3.5 多因子投资模型运行速度比较

本文通过计算各个部分计算时间来评价计算机性能。总体来说，除了分布式高效提升器以外，时间窗口越长，模型越高级，模型拟合和投资组合构造所需要的时间越长。

极端梯度提升树和分布式高效提升器各时间窗口平均的运行时间分别为2079.8 分钟、830.3 分钟，夏普比率分别为 2.30、2.31，夏普比比上运行时间（单位：百分比/分钟）分别为 0.0011、0.0028，计算效率提升将近 2.6 倍。尤其是分布式高效提升器算法，在策略表现差异不大的前提下，它的运行时间是最优的。

表 3.9 模型拟合及投资组合构造运算时间

算法	12 个月	36 个月	60 个月
FM	1:42:35	4:11:39	6:06:18
EXT	2:54:37	7:11:41	9:55:21
GBRT	4:29:28	9:44:48	15:56:08
XGB	10:15:25	7:41:42	16:42:42
light BGM	9:31:32	2:48:29	1:30:19

表 3.10 DM_test 运算时间

	EXT	GBRT	XGB	lightBGM
FM	1:10:18	4:09:02	5:17:10	0:49:26
EXT		14:07:33	5:32:52	1:17:35
GBRT			6:54:45	5:36:33
XGB				10:16:15

3.3.6 构造市值加权投资组合进行稳健性检验

单因子多空组合检验结果显示，市值因子（size）构建的 12 月滚动多空投资组合能够获得高达 21.26%的年化收益率；虽然小市值策略从历史上看表现优异，但是随着中国资本市场的进一步发展，各行业发展进入平稳期，竞争越来越激烈，市场格局越来越稳定，小市值效应在未来存在失效的风险，小市值效应所带来的历史数据问题可能对实验结果存在影响。为了本文用市值加权的方法构造投资组合对 3.3.2 的内容进行实验。

结果表明，简单加权投资组合各模型绩效表现均优于市值加权投资组合的结果，其他结论与 3.33 多因子投资策略实证结果分析中的结论保持一致。

表 3.11 投资策略绩效表现分析（市值加权平均）

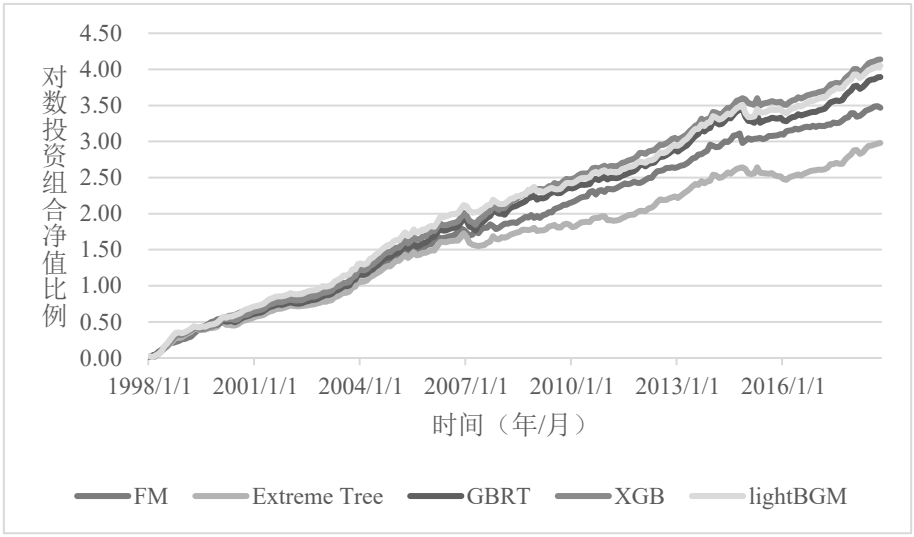
市值加权平均	12 个月					36 个月					60 个月				
模型	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM
多空组合年化收益率	0.4063	0.3543	0.4628	0.4883	0.4802	0.4947	0.3741	0.4555	0.4695	0.5066	0.3558	0.3634	0.4639	0.4745	0.5222
多空组合夏普比率	1.0579	0.7016	1.1234	1.2933	1.2070	1.5413	0.8974	1.2354	1.3319	1.4497	0.7508	0.7937	1.1902	1.2816	1.5374
多空组合 FF3-alpha	0.4143	0.3894	0.5043	0.5256	0.5163	0.5471	0.3862	0.5229	0.5371	0.5788	0.3535	0.4209	0.5242	0.5335	0.5794
多空组合 FF5-alpha	0.4359	0.3540	0.4584	0.4831	0.4732	0.4607	0.3860	0.4562	0.4632	0.5110	0.3730	0.3792	0.4826	0.4862	0.5249
多头组合年化收益率	0.2074	0.1648	0.2023	0.2171	0.2110	0.2777	0.1631	0.2129	0.2209	0.2540	0.1949	0.1682	0.2031	0.2188	0.2488
多头组合夏普比率	-0.0133	-0.1460	-0.0293	0.0185	-0.0012	0.2427	-0.0367	0.0429	0.0669	0.1644	-0.0197	-0.0989	0.0062	0.0535	0.1384
多头组合 FF3-alpha	0.2234	0.0860	0.1310	0.1439	0.1355	0.2071	0.1826	0.1543	0.1579	0.1872	0.1909	0.0937	0.1367	0.1494	0.1792
多头组合 FF5-alpha	0.2541	0.0802	0.1292	0.1409	0.1275	0.1809	0.1839	0.1431	0.1424	0.1698	0.2426	0.0923	0.1393	0.1501	0.1723
空头组合年化收益率	0.0124	0.0219	-0.0492	-0.0598	-0.0578	-0.0174	-0.0115	-0.0431	-0.0491	-0.0531	0.0402	0.0059	-0.0596	-0.0546	-0.0723
空头组合夏普比率	-0.6377	-0.5742	-0.7948	-0.8390	-0.8126	-0.6906	-0.0367	-0.7424	-0.7612	-0.7554	-0.5028	-0.5611	-0.7510	-0.7373	-0.7888
空头组合 FF3-alpha	0.0191	-0.0933	-0.1632	-0.1716	-0.1707	-0.1402	-0.0039	-0.1688	-0.1794	-0.1918	0.0388	-0.1257	-0.1860	-0.1826	-0.1987
空头组合 FF5-alpha	0.0286	-0.0634	-0.1187	-0.1318	-0.1353	-0.0804	-0.0027	-0.1137	-0.1214	-0.1418	0.0704	-0.0861	-0.1425	-0.1353	-0.1518

表 3.12 投资组合绩效指标显著性（市值加权平均）

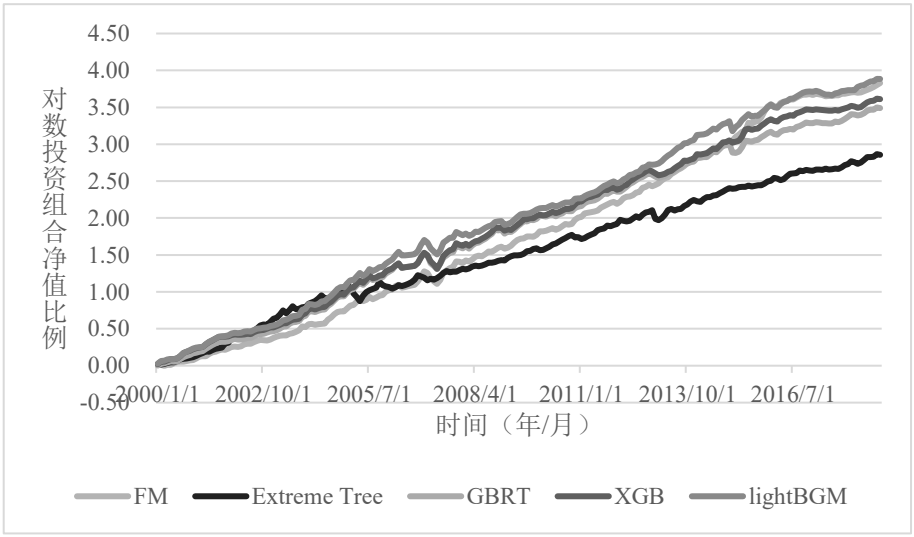
市值加权平均	12 个月					36 个月					60 个月				
模型	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM	FM	EXT	GBRT	XGB	LBGM
多空组合年化收益率	34.9336	27.5506	34.9336	34.9336	34.1578	38.7443	28.1936	32.9718	34.7359	35.8712	24.4876	25.1979	29.7902	31.5377	35.4502
多空组合 FF3-alpha	9.9577	8.8143	10.3002	11.3687	10.4485	13.3910	8.2245	11.6190	12.3484	13.1246	6.8945	9.2922	10.4752	11.1967	12.6695
多空组合 FF5-alpha	10.1300	7.9790	9.2883	10.4816	9.4008	11.1666	7.7525	9.6691	10.2691	11.1857	6.8267	7.8670	9.0147	9.6542	10.9929
多头组合年化收益率	10.9452	8.1870	10.9452	10.9452	10.5625	12.9229	7.6618	10.1987	10.3394	11.4927	8.7075	7.1638	8.8388	9.3827	10.2446
多头组合 FF3-alpha	3.3386	2.8842	4.0595	4.6053	4.1180	7.3873	2.4320	5.0674	5.2710	6.2136	2.4263	3.1005	4.0723	4.5460	5.8386
多头组合 FF5-alpha	3.6676	2.6286	3.8825	4.3841	3.7400	6.1938	2.3123	4.3993	4.4235	5.3350	2.9126	2.8742	3.8818	4.2685	5.2928
空头组合年化收益率	0.6322	1.0497	0.6322	0.6322	-2.7663	-0.8327	-0.5269	-1.9771	-2.2560	-2.3827	1.7815	0.2413	-2.4349	-2.2310	-2.9558
空头组合 FF3-alpha	0.2727	-3.5767	-6.0923	-6.5539	-6.4611	-5.4193	-0.0496	-6.0905	-6.6695	-7.1197	0.4868	-4.5914	-6.1286	-6.2134	-7.0794
空头组合 FF5-alpha	0.3914	-2.2985	-4.3623	-4.9699	-4.9436	-2.9020	-0.0323	-3.7863	-4.2022	-4.7854	0.8323	-2.6612	-4.1197	-4.0067	-4.7844

表 3.13　Boosting 集成算法策略结果分析（市值加权平均）

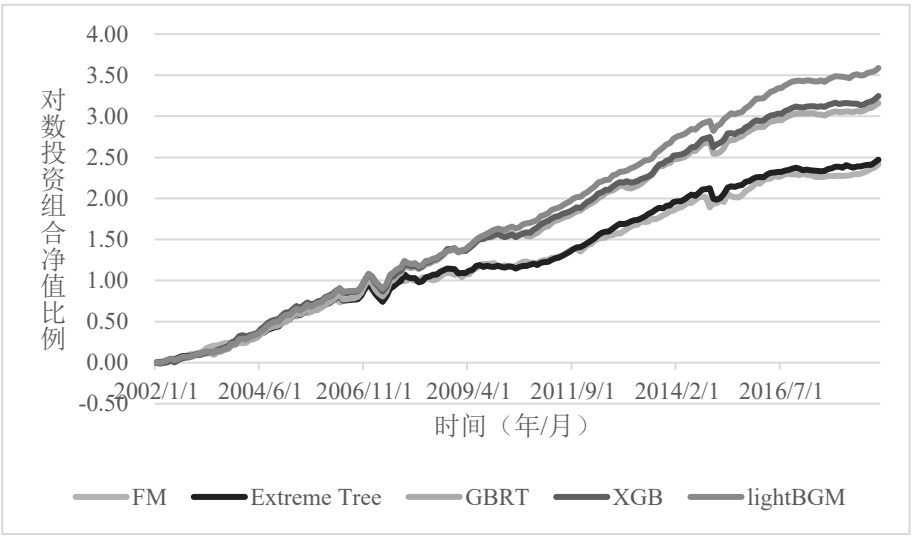
市值加权平均	12 个月			36 个月			60 个月			各时间窗口平均		
模型	FM	Boosting	变化情况	FM	Boosting	变化情况	FM	Boosting	变化情况	FM	Boosting	变化情况
多空组合年化收益率	0.4063	0.4464	0.0401	0.4947	0.4514	-0.0433	0.3558	0.4560	0.1002	0.4189	0.4513	0.0324
多空组合夏普比率	1.0579	1.0813	0.0234	1.5413	1.2286	-0.3127	0.7508	1.2007	0.4499	1.1167	1.1702	0.0535
多空组合 FF3-alpha	0.4143	0.4839	0.0696	0.5471	0.5063	-0.0408	0.3535	0.5145	0.1610	0.4383	0.5016	0.0632
多空组合 FF5-alpha	0.4359	0.4422	0.0063	0.4607	0.4541	-0.0066	0.3730	0.4682	0.0952	0.4232	0.4548	0.0316
多头组合年化收益率	0.2074	0.1988	-0.0086	0.2777	0.2127	-0.0650	0.1949	0.2097	0.0149	0.2267	0.2071	-0.0196
多头组合夏普比率	-0.0133	-0.0395	-0.0262	0.2427	0.0594	-0.1834	-0.0197	0.0248	0.0445	0.0699	0.0149	-0.0550
多头组合 FF3-alpha	0.2234	0.1241	-0.0993	0.2071	0.1705	-0.0366	0.1909	0.1397	-0.0511	0.2071	0.1448	-0.0623
多头组合 FF5-alpha	0.2541	0.1194	-0.1346	0.1809	0.1598	-0.0211	0.2426	0.1385	-0.1041	0.2259	0.1393	-0.0866
空头组合年化收益率	0.0124	-0.0362	-0.0487	-0.0174	-0.0392	-0.0218	0.0402	-0.0451	-0.0854	0.0117	-0.0402	-0.0519
空头组合夏普比率	-0.6377	-0.7552	-0.1174	-0.6906	-0.5739	0.1167	-0.5028	-0.7095	-0.2067	-0.6104	-0.6795	-0.0692
空头组合 FF3-alpha	0.0191	-0.1497	-0.1688	-0.1402	-0.1360	0.0043	0.0388	-0.1733	-0.2121	-0.0274	-0.1530	-0.1256
空头组合 FF5-alpha	0.0286	-0.1123	-0.1409	-0.0804	-0.0949	-0.0144	0.0704	-0.1289	-0.1993	0.0062	-0.1120	-0.1182



(a)时间窗口为 12 个月



(b)时间窗口为 36 个月



(c)时间窗口为 60 个月

图 3.3 投资组合对数净值比例图（市值加权）

4 结论与展望

4.1 结论

股票市场的量化多因子回归模型是资产定价领域的经典研究问题。机器学习算法在股票市场的量化多因子回归的运用能够高效识别有用的因子，解决高维特征等问题，能够更好地解释和预测股票市场的超额收益。为了验证现有文献中的因子在中国市场的适用性以及 Boosting 系列算法在投资学领域的适用性，本文通过使用 Boosting 系列机器学习算法包括极限随机森林、梯度提升决策树、极端梯度提升树、分布式高效提升器等，选取中国 A 股市场 1997 年-2018 年月度数据，构造 96 个单因子输入模型中，进行滚动窗口的回归预测，窗口期分别采取 12 个月、36 个月和 60 个月。

第一，从投资组合构造的角度上看，多因子模型的效果显著优于单因子模型；Boosting 机器学习多因子模型的效果优于传统的 Fama Macbeth 简单线性多因子模型。在简单加权构造的投资组合中，60 个月的分布式高效提升器算法多空投资组合的夏普比率可以达到 2.52、年化收益率可达到 64.63%，多空投资组合的年化 FF3-Alpha 收益可以达到 57.94%。

第二，从特征筛选的角度上看，无论是考虑单因子检验还是考虑多因子模型特征重要性评价指标的结果，交易摩擦类因子和动量类因子在中国 A 股市场占主导地位，但 Boosting 算法能更好地识别财务流动性因子。综合来看，交易摩擦类因子、动量因子和财务流动性因子在中国 A 股市场重要性排前三位，在考虑了多因子的共同作用的情况下，中国市场有其内在的市场有效性。

第三，机器学习回归算法与 Fama Macbeth 算法存在差异，但是各个 Boosting 算法的效果在迪堡马里亚诺检验上不存在显著差异。从计算机性能上看，分布式高效提升器在各个 Boosting 算法中占有极强的优势，计算效率可以提高近 3 倍。

4.2 展望

本文还有诸多不足，希望能在未来的研究中加以改进：

第一，本文未考虑交易成本等模型实际运用过程中不可忽略的因素（DeMiguel, etal, 2019）[48] 。在实际交易过程中，每次交易都会产生单边的交易手续费，不考虑交易成本即假定市场是不存在摩擦的完美市场，不符合现实，使模型与实际情

况脱轨，投资策略无法直接运用到实盘交易中。考虑交易成本时，可能存在模型失效的情况，需要后续研究来加以验证。

第二，市值因子对实证结果存在重大影响。各多因子模型对应的市值加权和简单加权两种方式构造的投资策略结果虽然从数据趋势和模型表现上看保持一致，但是从相对大小上看存在一定的差异。虽然小市值策略在历史上表现优异，但是随着中国资本市场的发展，市场格局趋于稳定，小公司“激流勇进”的概率越来越小，小市值效应在未来可能失效。未来的改进方向主要有两个：一是不采用 A 股全样本进行建模，去掉小市值股票，再次进行检验；二是去掉和市值有关的因子进行检验，以此来补充和拓展本文的稳健性检验的部分。

第三，Boosting 算法模型直接应用在投资领域存在局限性。首先，各 Boosting 模型对虚拟变量比较敏感，它们在因子重要性描述的过程中无法将虚拟变量和连续型变量一同进行比较，使得本文研究的多因子模型在因子重要性描述这一部分存在缺憾；其次，滚动调参的 Boosting 算法对计算机性能的要求在当前难以满足，未来在计算性能甚至在模型框架上存在一定的优化空间，可以实现在每月的截面预测回归模型上实现超参数的自动调整。

第四，本文所使用的滚动时间窗口内的月度截面模型集成算法以及 Boosting 模型集成算法均为简单的平均集成结果，存在改进空间。例如，对于一个滚动时间窗口内的月度截面模型，可以根据其数据的周期性、波动性或者通过更高的机器学习模式识别算法、识别股票市场趋势（牛市或熊市）来对每个月的模型赋予差异化的权重；对于模型集成算法，可以根据每个模型在每个月内表现的不同来在月度上赋予不同的权重来集成。尽管在当前简单的集成方式下，本文的各模型算法均有优异的表现，在改进的集成算法下模型表现会有较大的提升。

第五，本文仅使用 96 个因子，在未来拟使用更多的因子来构造特征工程，适应更高级的 Boosting 算法，比如按特征分类梯度提升决策树（CatBoost）。按特征分类梯度提升决策树能解决原始梯度提升决策树中的各种数据偏移问题，但是需要预先构造特征工程。

参考文献

[1] Antti I, 钱磊译. 预期收益——投资者获利指南[M].2011, 2018 年 5 月第 1 版，上海，格致出版社，上海人民出版社，2018：62-166,

[2] Fama, E, F. and K,R.French. Common Risk Factors in the Returns on Stocks and Bonds[J]. Journal of Financial Economics,1993, 33(1): 3-56.

[3] Carhart, M. M, On Persistence in Mutual Fund Performance[J]. The Journal of Finance, 1997, 52(1): 57-82.

[4] Fama, E, F. and K,R.French., A five-factor asset pricing model[J]. Journal of Financial Economics,2015, 116(2015): 1-22.

[5] Fama, E, F. and K,R.French, International Tests of a Five- Factor Asset Pricing Model[R]. Tuck School of Business Working Paper, 2015, No.2622782.

[6] Fama, E, F. and K,R.French, Dissecting Anomalies with a Five-Factor Model[J]. Review of Financial Studies, 2016, 29(1): 69-103.

[7] Fama, E, F, MacBeth. J, D, Risk, Return, and Equilibrium: Empirical Tests[J]. Journal of Political Economy,1973, 81(3): 607-636.

[8] William F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk[J]. The Journal of Finance,1964, 19(3): 425-442.

[9] Geurts P, Ernst D, and Wehenkel L. Extremely randomized trees[J]. Machine Learning,2006, 63(1): 3-42.

[10] Friedman J. Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

[11] Friedman J. Stochastic Gradient Boosting[R]. 1999.

[12] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[R]. 2016, In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.

[13] Ridgeway, Generalized Boosted Models: A guide to the gbm package[R]. 2007.

[14] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree[J]. 2017, Advances in Neural Information Processing Systems,2017, 30: 3149-3157.

[15] Fama, E, F. and K,R.French. The Cross-Section of Expected Stock Returns[J]. The Journal Of Finance, 1992, 47,(2):427-465.

[16] Stambaugh, R.F., Yu, J. and Yuan, Y. The short of it: Investor sentiment and anomalies[J]. Journal of Financial Economics, 2012, 104(2): 288–302.

[17] Novy-Marx, R., The Other Side of Value: The Gross Profitability Premium[J]. Journal of Financial Economics, 2013, 108(1): 1-28.

[18] Lewellen, Jonathan W., The Cross Section of Expected Stock Returns[J] Critical Finance Review, Critical Finance Review, 2015, 4: 1–44.

[19] Green, J, Hand, J R. M. and Zhang X F, The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns[J]. The Review of Financial Studies, 2017, 30(12): 4389-4436.

[20] McLean, R.D. and Pontiff, J. Does academic research destroy stock return predictability? [J]. The Journal of Finance, 2016, 71(1): 5–32.

[21] Hou K. Xue C. Zhang L. Replicating Anomalies[J]. The Review of Financial Studies, 2019, forthcoming.

[22] Hu G X, Pan J, Wang J, Chinese Capital Market: An Empirical Overview[R]. NBER Working Paper No. w24346, 2018.

[23] 朱英伦,刘杰.中国股票市场因子研究综述[J].现代管理科学,2018(07):42-44.

[24] 李志冰、杨光艺、冯永昌、景亮. Fama－French 五因子模型在中国股票市场的实证检验[J], 金融研究. 2017, 6：191-206.

[25] Guo, B. W．Zhang, Y. J．Zhang and H. Zhang, The Five Factor Asset Pricing Model Tests for the Chinese Stock Market[J]. Pacific－Basin Finance Journal，2017, 43: 84-106．

[26] Jason. V. Michael .Anomalies in Chinese A-Shares[R].2017.

[27] 中国 A 股市场量化因子白皮书. [R].北京：清华大学五道口金融学院.2018.

[28] 朱英伦,刘杰.中国股票市场因子研究综述[J].现代管理科学,2018, 7: 42-44.

[29] Moritz B, Zimmermann T. Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns[J]. SSRN, 2016.

[30] Kozak S, Nagel S, Santosh S, Shrinking the Cross Section[J]. NBER Working Papers 24070, National Bureau of Economic Research, Inc. 2017.

[31] 谢合亮,胡迪.多因子量化模型在投资组合中的应用——基于 LASSO 与 Elastic Net 的比较研究[J].统计与信息论坛,2017,32(10): 36-42.

[32] Serhiy S Shrihari. Interpreting Factor Models[J].The Journal of Finance, 2018, 73(3): 1214-1228.

[33] Tsai C, Hsiao Y. Combining multiple feature selection methods for stock

prediction Union, intersection, and multi-intersection approaches[J]. Decision Support Systems, 2010, 50(1): 258-269.

[34] Feng. Polson._Xu. Deep Learning Factor Alpha[R]. Finance 2018, EcoSta 2018, SOFIE Summer School, 2018.

[35] Gu S, Kelly B T, Xiu D. Empirical Asset pricing via machine learning[J]. Chicago Booth Research Paper No. 18-04, 2018.

[36] Diebold, F. X, Mariano, R. S. Comparing predictive accuracy[J]. Journal of business & economic statistics, 1995, 13(3): 253-264.

[37] Bryan. Seth, Market Expectations in the Cross-Section of Present Values[J]. The Journal of Finance, 2013, 67(5): 1721-1756.

[38] Ondrej. T. Martin H., Does It Pay to Follow Anomalies Research? Machine Learning Approach with International Evidence[R]. 2018.

[39] Jiang F, Tang G, Zhou G. Firm Characteristics and Chinese Stocks[J]. Journal of Management Science and Engineering, 2018, 3(4): 259–284.

[40] Leo B, Random forests[J], Machine learning, 2001, 45: 5-32.

[41] Leo B, Friedman J, Charles J Stone, Richard A Olshen, Classification and regression trees[M], CRC press, 1984: 304-384

[42] Meir R, Rätsch G, An introduction to boosting and leveraging[M]. In S. Mendelson and A. Smola, editors, Advanced Lectures on Machine Learning, LNCS, pp 119-184. Springer, 2003. In press. Copyright by Springer Verlag.

[43] Meng Q, Ke Q, Wang T, Chen W, Ye Q, Ma Z, Liu T. A Communication-Efficient Parallel Algorithm for Decision Tree[J]. 2016, Advances in Neural Information Processing Systems, 2016, 29: 1279-1287.

[44] Huan Zhang, Si Si and Cho-Jui Hsieh. GPU Acceleration for Large-scale Tree Boosting[R]. SysML Conference, 2018.

[45] 吴林祥,徐龙炳,王新屏.价格涨跌幅限制起到了助涨助跌作用吗?[J].经济研究,2003(10):59-65+93.

[46] 金超,柯昌隆.基于信息经济学的 IPO 溢价研究综述[J].商业经济,2011(17):24-25+34.

[47] Low, R.K.Y.; Alcock, J.; Faff, R.; Brailsford, T. Canonical vine copulas in the context of modern portfolio management: Are they worth it?[J]. Journal of Banking & Finance. 2013, 37 (8): 3085–3099.

[48] DeMiguel, Victor and Martin-Utrera, Alberto and Nogales, Francisco J. and Uppal, Raman, A Transaction-Cost Perspective on the Multitude of Firm Characteristics[R]. 2019.

致谢

毕业论文即将进入尾声，我按捺不住激动的心情。在这六个月的实证研究、文献阅读和论文撰写过程中，我要感谢很多人。

首先，我要感谢我的指导老师李斌老师。本研究及学位论文是在李老师的亲切关怀和悉心指导下完成的。从论文选题、数据处理、到实证部分以及最后的论文写作，李老师始终悉心给予我指导和支持，我也被李老师严谨认真的治学态度和孜孜不倦、精益求精的精神所感动。在此谨向李老师致以诚挚的谢意和崇高的敬意！

其次，我要感谢中国 A 股市场量化异象因子数据库的所有成员，包括邵新月学姐、李玥阳学长、岳阳同学。有了你们技术和精神上的支持，我的论文才能有如此完整、庞大且准确的数据库作为支撑。其中，邵新月学姐和孙橙学姐面对我的“问题轰炸”仍能耐心细致地回答，对我的殷切支持也不断鼓励着我前行，真的非常感谢她们。

再次，我要感谢华中科技大学的谭重阳学长、香港中文大学（深圳）的佘胤娴同学以及四川大学的李沛同学。能借用他们实验室或者公司的服务器运行我冗长的代码为我的毕业论文实证部分结果的得出节约了很多时间，感谢他们在百忙之中给予我硬件上的支持；

最后，我要感谢我的父母、工十七 602 室全体室友和中国政法大学的雷钰文同学。对于从金融专业转到数据科学专业的我来说，量化研究是一条艰难的道路，这一路上我也时常迷茫、困惑。正是有了父母与朋友的陪伴、鼓励和支持，才得以让我坚持不懈地在量化道路上走下去！

附录

附录 A 因子数据说明

表 A.1 单因子简称、中文全称以及解释

	变量名	中文全称	英文
	A.交易摩擦因子(21 个)		
1	**size**	A 股流通市值	Firm size
2	**size_ia**	行业调整市值	Industry-adjusted size
3	**beta**	系统性风险	Market beta
4	**betasq**	系统性风险的平方	Market beta squared
5	**betad**	下行风险	Downside beta
6	**idvol**	异质波动率	Idiosyncratic volatility
7	**vol**	总波动率	Total volatility
8	**idskew**	特定偏态	Idiosyncratic skewness
9	**skew**	总偏态	Total skewness
10	**coskew**	共同偏态	Co-skewness
11	**turn**	交易换手率	Trading turnover
12	**std_turn**	换手率的波动率	Volatility of turnover
13	**volumed**	交易额	Volume in dollar
14	**std_dvol**	交易额的波动率	Volatility of volume in dollar
15	**retnmax**	最大日收益率	Maximum daily return
16	**illq**	非流动性风险	Illiquidity
17	**LM**	标准化的换手率	Zero trade
18	**sharechg**	股本的增长率	Anuual percent Changes in share
19	**age**	公司年龄	Firm age since IPO
20	**aeavol**	收益公告异常交易量	Abnormal earnings announcement volume
21	**IPO**	新股发行	New equity issue
	B.动量因子(6 个)		
22	**mom12**	12 个月动量	12-month momentum
23	**mom6**	6 个月动量	6-month momentum
24	**mom36**	36 个月动量	36-month momentum
25	**momchg**	动量变化	Momentum change
26	**imom**	特定动量	Idiosyncratic momentum
27	**lagretn**	短期反转	Lagged return or reversal
	C.价值因子(10 个)		
28	**BM**	公司账面市值比	Book-to-market ratio
29	**BM_ia**	行业调整账面市值比	Industry adjusted book-to-market ratio
30	**AM**	总资产市值比	Asset-to-market ratio
31	**LEV**	总负债市值比	Leverage
32	**EP**	收益价格比	Earnings to price ratio
33	**CFP**	现金流价格比	Cashflow to price ratio
34	**CFP_ia**	行业调整 CFP	Industry adjusted CFP
35	**OCFP**	营业现金流价格比	Operating cashflow to price ratio
36	**DP**	股利价格比	Dividend to price ratio
37	**SP**	营业收入价格比	Sales to price ratio

（续表 A.1 单因子简称、中文全称以及解释）

	D.成长因子(29 个)		
38	**AG**	总资产增长率	Asset growth
39	**LG**	总负债增长率	Liabilities growth
40	**SG**	营业收入增长率	Sales growth
41	**PMG**	营业利润增长率	Profit margin growth
42	**INVG**	存货增长率	Inventory growth
43	**INVchg**	存货变化	Inventory change
44	**SgINVg**	营业收入与存货增长率的差	Sales growth minus inventory growth
45	**TAXchg**	税收增长率	Tax change
46	**ACC**	应记项目	Accruals
47	**absacc**	应计项目绝对值	Absolute accruals
48	**stdacc**	应计项目波动率	Accrual volatility
49	**ACCP**	应计项目变化	Percent accruals
50	**cinvest**	公司投资	Corporate investment
51	**depr**	折旧率	Depreciation/PP&E
52	**pchdepr**	折旧变动百分比	%change in depreciation
53	**egr**	股东权益变化	Change in shareholders’equity
54	**fgr5yr**	预期 5 年每股收益增长	Forecasted growth in 5-year EPS
55	**grCAPX**	资本支出变化	Percent change in capital expenditures
56	**pchcapx_ia**	行业调整后资本支出变动百分比	Industry-adjusted% change in capital expenditures
57	**grltnoa**	长期净经营资产增长	Growth in long-term net operating assets
58	**invest**	资本支出和存货变化	Capital expenditures and inventory
59	**pchsale_pchinvt**	销售增长减存货增长	%change in sales-%change in inventory
60	**pchsale_pchrect**	销售增长减应收账款增长	%change in sales-%change in A/R
61	**pchsale_pchxsga**	销售增长减 SG&A 增长	%change in sales-%change in SG&A
62	**realestate**	不动产持有量	Real estate holdings
63	**sgr**	销售增长	Sales growth
64	**NOA**	净经营资产	Net operating assets
65	**hire**	雇员增长率	Employee growth rate
66	**chempia**	行业调整后雇员人数变动	Change in number of employees

（续表 A.1 单因子简称、中文全称以及解释）

	E.盈利因子(18 个)		
67	**ROE**	净资产收益率	Return on equity
68	**ROA**	总资产收益率	Return on asset
69	**CT**	资本换手率	Capital turnover
70	**PA**	利润资产比率	Profit-to-assets
71	**cashpr**	现金生产力	Cash productivity
72	**cash**	现金净资产比	Cash
73	**RD**	研发成本	Research and development(R&D)
74	**rd_mve**	研发支出市值比	R&D to market capitalization
75	**RDsale**	研发成本收入比	R&D to sales ratio
76	**operprof**	营业利润率	Operating profit rate
77	**pchgm_pchsale**	毛利率变动%-销售变动%	%change in gross margin-% change in sales
78	**ATO**	总资产周转率	Asset turnover
79	**chfeps**	预期每股收益的变化	Change inforecasted EPS
80	**nincr**	收入增加期数	Number of earnings increases
81	**roic**	投入资本回报	Return on invested capital
82	**rsup**	意外收入	Revenue surprise
83	**sue**	未预期收益	Unexpected quarterly earnings
84	**sfe**	收益预测	Scaled earnings forecast
	F.财务流动性因子(10 个）		
85	**CR**	流动比率	Current ratio
86	**QR**	速动比率	Quick ratio
87	**CFdebt**	现金流负债比	Cashflow to debt ratio
88	**salecash**	营业收入现金比	Sales to cash ratio
89	**saleinv**	营业收入存货比	Sales to inventory
90	**CRG**	流动比率增长率	Current ratio growth
91	**QRG**	速动比率增长率	Quick ratio growth
92	**pchsaleinv**	营业收入存货比增长率	%change sales-to-inventory
93	**salerec**	营业收入应收账款比	Sales to receivables
94	**tang**	偿债能力/总资产	Debt capacity/firm tangibility
	H.其他（6 个）		
95	**chnanalyst**	分析师人数变化	Change in number of analysts
96	**nanalyst**	涉及股票的分析师人数	Number of analysts covering stock
97	**divi**	发放股利指标	Dividend initiation
98	**divo**	停止股利指标	Dividend omission
99	**herf**	行业销售集中度	Industry sales concentration
100	**sin**	有罪的股票	Sin stocks

附录 B 因子数据集描述性统计分析

表 B.1 因子描述性统计分析

A.交易摩擦因子(21 个)									
	因子	count	mean	std	min	25%	50%	75%	max
1	size	1679	4.27E+06	1.54E+07	1.92E+05	1.18E+06	2.02E+06	3.81E+06	5.82E+08
2	size_ia	1679	0	1.51E+07	-2.02E+07	-2.61E+06	-1.66E+06	-2.10E+04	5.62E+08
3	beta	1440	1	0	0	1	1	1	3.04
4	betasq	1440	1	2	0	1	1	1	48.33
5	betad	1440	1	0	0	1	1	1	3.59
6	idvol	1440	1	0	1	1	1	2	1.91
7	vol	1618	3	2	1	2	3	3	41.02
8	idskew	1440	-1	-1	-13	-1	-1	0	2.77
9	skew	1440	0	1	-2	0	0	0	12.28
10	coskew	1440	0	5	-17	-3	-1	2	60.99
11	turn	1446	2	1	0	2	2	3	10.75
12	std_turn	1643	1	2	0	1	1	2	18.39
13	volumed	1446	7.84E+07	9.41E+07	6.05E+06	3.17E+07	5.10E+07	8.83E+07	1.40E+09
14	std_dvol	1643	3.97E+07	6.06E+07	1.24E+06	1.23E+07	2.26E+07	4.41E+07	1.08E+09
15	retnmax	1636	0	0	0	0	0	0	2.40
16	illq	1632	0	0	0	0	0	0	0.19
17	LM	1643	3.97E+07	6.06E+07	1.24E+06	1.23E+07	2.26E+07	4.41E+07	1.08E+09
18	sharechg	1508	0	1	0	0	0	0	8.77
19	age	1449	7	4	0	4	7	11	16.54
20	aeavol	1425	0	7	-1	0	0	0	295.50
21	pricedelay	1689	1	0	0	1	1	1	1.00

B.动量因子(6 个)									
	因子	count	mean	std	min	25%	50%	75%	max
22	mom12	1719	1	1	0	1	1	1	13.67
23	mom6	1716	1146	1009	0	10	1135	2268	2289.10
24	mom36	1599	2	2	0	1	1	2	23.34
25	momchg	1647	0	1	-10	0	0	0	2.51
26	imom	1601	0	3	-26	-1	0	2	26.16
27	lagretn	1685	1146	1008	1	10	1135	2268	2283.87

（续表 B.1 因子描述性统计分析）

C.价值因子(10 个)									
	因子	count	mean	std	min	25%	50%	75%	max
28	BM	1630	361	249	-1604	226	333	469	2404.84
29	BM_ia	1630	0	240	-1942	-125	-21	107	2048.39
30	AM	1630	2041	2482	74	932	1446	2293	49061.57
31	LEV	1630	1097	1781	7	321	637	1237	37057.32
32	EP	1632	37	141	-2043	15	38	69	1373.05
33	CFP	1680	13	198	-1631	-53	-3	50	1982.35

（续表 B.1 因子描述性统计分析）

续 C.价值因子(10 个)									
	因子	count	mean	std	min	25%	50%	75%	max
34	CFP_ia	1680	0	196	-1631	-69	-16	41	1945.42
35	OCFP	1680	49	208	-1899	-13	32	93	2960.50
36	DP	1451	8	23	-2	0	1	8	401.34
37	SP	1630	1415	2932	-8	376	757	1511	60829.00

D.成长因子(29 个)									
	因子	count	mean	std	min	25%	50%	75%	max
38	AG	1555	1	30	-1	0	0	0	1218.82
39	LG	1555	24	1163	-31	0	0	0	56786.09
40	BVEG	1555	-1	95	-4211	0	0	0	392.64
41	SG	1552	4	154	-191	0	0	0	6862.89
42	PMG	1558	2	103	-1157	0	0	0	2576.80
43	INVG	1520	18	796	-1	0	0	0	36322.59
44	INVchg	1519	0	0	-1	0	0	0	0.93
45	SgINVg	1518	-15	801	-34157	0	0	0	3348.05
46	TAXchg	1544	8	497	-3496	-1	0	1	18831.04
47	ACC	1847	0	6	-251	0	0	0	43.77
48	absacc	1842	0	6	0	0	0	0	257.91
49	stdacc	1909	0	4	0	0	0	0	181.88
50	ACCP	1710	-9484	377817	-1.45E+07	-1	0	2	182902.10
51	cinvest	1793	28	978	-96	0	1	1	36174.75
52	depr	1627	2	68	0	0	0	0	2967.54
53	pchdepr	1684	2	28	-37	1	1	1	1211.49
54	egr	1649	1	15	-204	1	1	1	441.38
55	fgr5yr	1310	25	86	-3	2	7	23	3046.45
56	grCAPX	1688	28	1016	-1	1	1	2	47704.24
57	pchcapx_ia	1676	0	997	-644	-3	-3	-1	46473.61
58	grltnoa	1650	1	14	-239	1	1	1	437.19
59	invest	1688	36	1192	-2	2	2	3	54015.12
60	pchsale_pchinvt	1571	-2	330	-7440	0	0	0	6852.98
61	pchsale_pchrect	1558	-6	564	-22347	0	0	0	6756.76
62	pchsale_pchxsga	1559	3	157	-322	0	0	0	6859.91
63	realestate	1699	1.94E+09	1.21E+10	-4.84E+06	1.50E+08	3.56E+08	9.23E+08	3.60E+11
64	sgr	1559	5	154	-190	1	1	1	6864.07
65	NOA	1701	3.43E+09	1.95E+10	-7.18E+09	5.11E+08	1.05E+09	2.27E+09	7.32E+11
66	hire	1614	3154	14038	20	689	1362	2642	404579.49
67	chempia	1610	0	13800	-13952	-2062	-1176	-11	390679.20

（续表 B.1 因子描述性统计分析）

E.盈利因子(18 个)

	因子	count	mean	std	min	25%	50%	75%	max
68	ROE	1561	20624	358760	-436741	0	0	0	1.17E+07
69	ROA	1562	19698	355917	-1125376	0	0	0	1.08E+07
70	CT	1700	3.77E+09	2.84E+10	-2.54E+06	3.14E+08	7.48E+08	1.94E+09	9.75E+11
71	PA	1701	22351	512593	-777351	0	0	0	1.74E+07
72	cashpr	1667	-34	396	-12783	-14	-8	-5	40.50
73	cash	1694	8.94E+08	3.31E+09	9.72E+04	1.19E+08	2.78E+08	6.34E+08	8.13E+10
74	RD	1693	1.80E+08	1.04E+09	-2.22E+07	2.93E+07	5.66E+07	1.21E+08	3.39E+10
75	rd_mve	1635	62	93	-31	22	39	69	1587.52
76	RDsale	1690	4	139	-173	0	0	0	4981.25
77	operprof	1681	0	2	-5	0	0	0	38.36
78	pchgm_pchsale	1532	-4	141	-4945	0	0	0	470.26
79	ATO	1682	1	23	-542	0	1	1	531.20
80	chfeps	3571	0	0	-4	0	0	0	5.18
81	nincr	3571	1	2	0	0	0	1	6.01
82	roic	1682	0	24	-56	0	0	0	1108.14
83	rusp	1620	67	3276	-57531	-9	64	213	20010.34
84	sfe	1865	3.04E+09	5.46E+11	-1.28E+13	-6.89E+08	0.00E+00	1.32E+09	1.27E+13

F.财务流动性因子(10 个)

	因子	count	mean	std	min	25%	50%	75%	max
85	CR	1682	2	7	-2	1	1	2	250.49
86	QR	1682	2	6	-1	1	1	2	229.63
87	CFdebt	1688	0	0	-5	0	0	0	8.19
88	salecash	1674	14	175	-1	2	3	7	4220.12
89	saleinv	1653	109	3284	-7	2	4	7	1.46E+05
90	CRG	1542	0	4	-3	0	0	0	124.09
91	QRG	1542	0	3	-2	0	0	0	100.29
92	pchsaleinv	1502	5497	252420	-1755	0	0	0	1.16E+07
93	salerec	1662	1340	60797	-32	2	4	9	2.94E+06
94	tang	1682	1.93E+09	7.90E+09	2.67E+06	3.09E+08	6.13E+08	1.32E+09	2.00E+11

H.其他（6 个）

	因子	count	mean	std	min	25%	50%	75%	max
95	chnanalyst	3571	0	7	-61	-1	0	1	65.33
96	nanalyst	3571	2	6	0	0	0	1	76.45
97	divi	-	-	-	-	-	-	-	-
98	divo	-	-	-	-	-	-	-	-
99	herf	-	-	-	-	-	-	-	-
100	sin	-	-	-	-	-	-	-	-

附录 C 各算法因子重要性程度

表 C.1 各算法的因子重要性程度（详细）

因子	FM-12	FM-36	FM-60	EXT-12	EXT-36	EXT-60	GBRT-12	GBRT-36	GBRT-60	XGB-12	XGB-36	XGB-60	lBGM-12	lBGM-36	lBGM-60
size	3.3055	3.0893	3.2461	0.1365	0.0981	0.1268	0.5473	0.0981	0.4586	0.2665	0.2925	0.2967	0.3767	0.4705	0.4601
size_ia	0.2580	0.0231	0.0014	0.0656	0.0554	0.0775	0.1515	0.0554	0.1958	0.2492	0.2487	0.2516	0.1593	0.2334	0.1990
beta	3.1607	0.7991	0.9738	0.2413	0.2458	0.2768	0.3272	0.2458	0.5349	0.4991	0.5697	0.6999	0.7817	0.7803	0.7386
betasq	3.4821	1.2795	1.4646	0.2026	0.2258	0.3383	0.4658	0.2258	0.5551	1.1273	1.1698	1.2057	2.4095	2.4142	2.4211
betad	1.0569	1.0617	1.2324	0.1155	0.1636	0.1279	0.4289	0.1636	0.4558	0.1678	0.1640	0.2222	0.6684	0.6243	0.5688
idvol	0.5622	0.6947	0.5746	0.0356	0.0414	0.0489	0.2524	0.0414	0.1675	0.0359	0.0303	0.0516	0.8777	0.8314	0.7351
vol	1.4298	1.4408	1.5915	0.2260	0.2509	0.2342	0.7652	0.2509	0.7579	0.1116	0.1420	0.1713	1.3966	1.3665	1.3040
idskew	0.0562	0.0495	0.0067	0.0525	0.0529	0.0440	0.0897	0.0529	0.1004	0.2311	0.2323	0.2338	0.6953	0.6423	0.6061
skew	0.0962	0.0415	0.0566	0.0819	0.0641	0.0599	0.2060	0.0641	0.2024	0.2669	0.2732	0.2734	0.5929	0.5535	0.4982
coskew	0.1887	0.2313	0.2856	0.0593	0.0629	0.0528	0.1345	0.0629	0.0801	0.1971	0.2339	0.2449	1.1114	1.0643	0.9974
Turn	0.5245	0.6120	0.7186	0.0125	0.0263	0.0038	0.0950	0.0263	0.0730	0.0878	0.0840	0.0659	0.6697	0.6081	0.5390
Std_turn	4.2311	4.1082	4.3285	0.0647	0.0644	0.0556	0.3292	0.0644	0.3465	0.0466	0.0537	0.0773	0.8201	0.7599	0.7657
Volumed	0.7066	1.0582	1.4396	0.0479	0.0143	0.0509	0.3030	0.0143	0.0795	0.4019	0.3763	0.3190	0.0284	0.0974	0.2182
std_dvol	0.5577	0.6199	0.8765	0.1706	0.1646	0.2162	0.7215	0.1646	0.7134	0.6146	0.6615	0.5770	0.2606	0.2294	0.2163
retnmax	0.1041	0.1513	0.1880	0.0313	0.0675	0.0151	0.0855	0.0675	0.0657	0.0226	0.0510	0.0612	0.9840	0.9325	0.8766
illq	0.5895	0.3587	0.1605	0.0034	0.0070	0.0194	0.3077	0.0070	0.1272	0.3964	0.3273	0.2447	0.1905	0.1308	0.0332
LM	0.6148	0.5470	0.6248	0.2037	0.2489	0.2459	0.4935	0.2489	0.5170	0.3500	0.3463	0.3673	0.4040	0.3559	0.3026
sharechg	0.0652	0.0618	0.0870	0.1840	0.1579	0.1429	0.1016	0.1579	0.1184	0.1076	0.1318	0.1395	0.6185	0.6736	0.7460
age	0.2696	0.1133	0.0961	1.0817	1.1629	1.1724	4.2822	1.1629	4.6759	1.8937	1.9970	2.1068	1.0435	1.1890	1.2927
aeavol	1.3611	1.2128	1.3022	1.0877	1.1686	1.1932	3.5011	1.1686	3.9156	0.5262	0.5881	0.6812	3.9751	4.0816	4.1881
pricedelay	0.0030	0.0184	0.0233	0.1166	0.1100	0.1220	0.2882	0.1100	0.2985	0.3656	0.3837	0.3868	0.9337	0.8607	0.8476
mom12	0.8079	0.9169	1.1763	0.4471	0.4262	0.4700	1.3125	0.4262	1.2280	0.4556	0.4781	0.4908	1.4161	1.3693	1.4072
mom6	1.2168	1.0945	1.2536	0.2558	0.2939	0.3102	0.7841	0.2939	0.8408	0.3021	0.3458	0.3289	1.2980	1.2749	1.2770
mom36	0.5390	0.5086	0.5360	0.0438	0.0224	0.0137	0.3155	0.0224	0.2625	0.0072	0.0028	0.0001	1.1632	1.1069	1.0950

（续表表 C.1 各算法的因子重要性程度（详细））

momchg	1.4413	1.5926	1.7429	0.2795	0.2701	0.2399	0.7337	0.2701	0.5823	0.1499	0.1386	0.1195	1.8898	1.7413	1.6585
imom	0.0257	0.0937	0.0796	0.0309	0.0298	0.0200	0.1844	0.0298	0.1204	0.2640	0.2671	0.2550	0.6962	0.6468	0.6959
lagretn	0.5836	0.8118	0.9276	0.6054	0.6576	0.6317	2.0736	0.6576	2.1300	0.3490	0.3748	0.3975	3.0530	2.9907	2.9702
BM	0.6896	0.6076	0.6028	0.0029	0.0311	0.0151	0.1751	0.0311	0.0674	0.2412	0.2320	0.2307	0.1113	0.1295	0.1805
bm_ia	0.6934	0.5610	0.4762	0.1166	0.0801	0.1216	0.3179	0.0801	0.3813	0.0170	0.0286	0.0513	0.2237	0.2482	0.2880
AM	0.2604	0.1279	0.0458	0.2388	0.2749	0.2662	0.4495	0.2749	0.4819	0.0934	0.1167	0.1242	0.3402	0.3472	0.3837
LEV	0.4000	0.0702	0.0425	0.2399	0.2790	0.2741	0.4844	0.2790	0.4613	0.3308	0.3336	0.3237	0.4756	0.4832	0.4735
EP	0.3473	0.2090	0.2224	0.1560	0.1446	0.1576	0.0634	0.1446	0.0831	0.1498	0.1769	0.1996	0.1712	0.2609	0.2876
CFP	0.4612	0.5867	0.5439	0.2705	0.2495	0.2453	0.6075	0.2495	0.5681	0.4219	0.3893	0.4012	0.5649	0.4497	0.4095
CFP_ia	0.6402	0.6908	0.7046	0.2715	0.2521	0.2599	0.6056	0.2521	0.5975	0.4484	0.4173	0.3990	0.6065	0.5050	0.4542
OCFP	0.2152	0.1854	0.1757	0.2421	0.2136	0.2356	0.4374	0.2136	0.4027	0.3419	0.3199	0.3114	0.0519	0.1696	0.2061
DP	0.0185	0.0287	0.0196	0.2419	0.2643	0.2922	0.5831	0.2643	0.6376	0.4609	0.4631	0.5316	0.6612	0.6850	0.7754
SP	0.4763	0.3807	0.4656	0.3109	0.2979	0.3129	0.5944	0.2979	0.5812	0.3227	0.3290	0.3351	0.4637	0.4747	0.4421
AG	0.2144	0.0768	0.2068	0.2061	0.2295	0.2337	0.5216	0.2295	0.5166	0.2324	0.2591	0.2501	0.1123	0.1242	0.1319
LG	0.0481	0.1478	0.2604	0.2237	0.2032	0.2229	0.4603	0.2032	0.4893	0.3006	0.3480	0.3535	0.2872	0.2534	0.2712
BVEG	0.9270	0.8658	0.8481	0.1730	0.1889	0.2118	0.3755	0.1889	0.4003	0.1836	0.1984	0.2141	0.0724	0.1016	0.0916
SG	0.4635	0.3717	0.1301	0.2721	0.3177	0.3114	0.6498	0.3177	0.6370	0.2719	0.2542	0.2754	0.4919	0.5282	0.5123
PMG	0.2616	0.1518	0.1702	0.0567	0.0707	0.0770	0.0389	0.0707	0.0356	0.0401	0.0484	0.0513	0.9382	1.0160	1.0299
INVG	0.2290	0.3151	0.3387	0.2241	0.2618	0.2852	0.5080	0.2618	0.5305	0.3419	0.3632	0.3862	0.2448	0.2211	0.2307
INVchg	0.4116	0.4182	0.5132	0.1798	0.1887	0.1730	0.3358	0.1887	0.3248	0.2618	0.2532	0.2494	0.0420	0.1090	0.1465
SgINVg	0.0140	0.0061	0.0081	0.3167	0.3572	0.3548	0.5889	0.3572	0.6146	0.4270	0.4318	0.4565	0.6260	0.6015	0.6091
TAXchg	0.1941	0.1846	0.2089	0.0549	0.1141	0.0987	0.2012	0.1141	0.2516	0.3036	0.3273	0.3447	1.1242	1.1675	1.1197
ACC	0.0857	0.1081	0.1233	0.2690	0.2449	0.2589	0.5514	0.2449	0.5584	0.3734	0.3413	0.3627	0.1481	0.0862	0.1049
absacc	0.2689	0.2525	0.2823	0.2465	0.2315	0.2041	0.4713	0.2315	0.4638	0.4133	0.3868	0.4020	0.0262	0.0413	0.0218
stdacc	0.0355	0.0078	0.0214	0.1527	0.1123	0.0866	0.3254	0.1123	0.2755	0.2658	0.1681	0.1505	0.1488	0.0931	0.1512
ACCP	0.0751	0.0621	0.0161	0.1945	0.1868	0.2089	0.4426	0.1868	0.4050	0.4623	0.4470	0.4572	0.2782	0.4228	0.4316
cinvest	0.1821	0.0999	0.0601	0.2636	0.2572	0.2583	0.4113	0.2572	0.4006	0.3123	0.2812	0.2889	0.0353	0.0928	0.1186

（续表表 C.1 各算法的因子重要性程度（详细））

depr	0.0919	0.0649	0.0295	0.2586	0.2441	0.2714	0.4413	0.2441	0.4335	0.4312	0.4127	0.4371	0.2228	0.3813	0.4382
pchdepr	0.0382	0.1351	0.0500	0.1777	0.1463	0.1759	0.4005	0.1463	0.3761	0.4573	0.4067	0.4322	0.4607	0.7120	0.7743
egr	2.6373	2.2505	2.2269	0.0044	0.0095	0.0075	0.5453	0.0095	0.5212	0.1180	0.1082	0.1421	1.0435	0.9589	0.9707
fgr5yr	1.7019	1.7036	1.7759	0.0095	0.0959	0.2087	0.4158	0.0959	0.7086	0.1509	0.0387	0.0797	0.2787	0.0499	0.2202
grCAPX	0.3142	0.2784	0.3045	0.3291	0.2711	0.2442	0.6135	0.2711	0.5807	0.4930	0.4241	0.4168	0.5498	0.3482	0.3084
pchcapx_ia	0.0640	0.1099	0.1315	0.2614	0.2153	0.2059	0.3272	0.2153	0.2325	0.3683	0.2836	0.2589	0.0873	0.1607	0.1904
grithnoa	0.1071	0.1809	0.1802	0.1901	0.1966	0.1942	0.3888	0.1966	0.4667	0.3148	0.3558	0.4215	0.6588	0.6317	0.6088
invest	0.2958	0.3480	0.3159	0.3442	0.3121	0.3010	0.5839	0.3121	0.5155	0.4771	0.4062	0.4119	0.5137	0.3084	0.2903
pchsale_pchinvt	0.0581	0.0414	0.0744	0.2634	0.2997	0.2980	0.5216	0.2997	0.5718	0.4587	0.4942	0.5287	0.1584	0.3060	0.3582
pchsale_pchrect	0.2764	0.1924	0.2593	0.2113	0.2189	0.2440	0.3724	0.2189	0.3953	0.2489	0.2534	0.2670	0.5141	0.5239	0.5468
Pchsale_pchxsga	0.2590	0.3121	0.3977	0.2787	0.2904	0.3186	0.5461	0.2904	0.5828	0.3742	0.4038	0.4349	0.1408	0.1746	0.2123
Realestate	0.3472	0.2558	0.3644	0.1140	0.0955	0.1325	0.0690	0.0955	0.0719	0.1506	0.1408	0.1770	0.2077	0.3014	0.3084
Sgr	0.1631	0.1690	0.1854	0.3076	0.3416	0.3544	0.6997	0.3416	0.6826	0.8270	0.8630	0.8717	1.7914	1.7762	1.7624
NOA	0.4193	0.4928	0.5173	0.0249	0.0508	0.0554	0.0572	0.0508	0.0610	0.3589	0.3618	0.4324	0.3779	0.4430	0.4429
hire	0.0463	0.1452	0.0254	0.1470	0.1910	0.1919	0.3467	0.1910	0.4333	0.2846	0.3047	0.3482	0.2725	0.2375	0.1739
chemp_ia	0.2198	0.1735	0.2203	0.1588	0.2128	0.1959	0.2237	0.2128	0.2530	0.2547	0.3013	0.3144	0.3246	0.2375	0.2073
ROE	0.0548	0.1571	0.2165	0.1718	0.1826	0.1941	0.3441	0.1826	0.4319	0.1452	0.0568	0.0265	0.6849	0.6706	0.6829
ROA	0.2186	0.0303	0.0864	0.1656	0.1844	0.2026	0.3663	0.1844	0.4060	0.1779	0.1423	0.1496	1.0165	1.0013	1.0122
CT	0.3876	0.2387	0.0631	0.2070	0.2120	0.1930	0.4607	0.2120	0.4159	0.1278	0.1751	0.1665	0.7592	0.8234	0.8209
PA	0.1862	0.2303	0.0765	0.2548	0.2434	0.2488	0.4821	0.2434	0.5278	0.0626	0.1733	0.1969	1.1116	1.1563	1.1487
cashpr	0.2982	0.2756	0.2984	0.2540	0.2481	0.2416	0.4839	0.2481	0.4713	0.2505	0.2523	0.2534	0.1869	0.1513	0.1217
cash	0.1909	0.1755	0.0755	0.1964	0.2338	0.2314	0.5168	0.2338	0.5237	0.1645	0.1837	0.1765	0.4650	0.5666	0.5699
RD	1.2367	0.9785	1.0069	0.2520	0.2851	0.2591	0.4469	0.2851	0.5070	0.2244	0.2685	0.2626	0.3075	0.3662	0.3586
rd_mve	2.1369	1.8324	1.9758	0.2529	0.2619	0.2291	0.5112	0.2619	0.4986	0.2915	0.3094	0.3486	0.1022	0.1026	0.0983
RDsale	0.2353	0.2435	0.3867	0.2092	0.2349	0.2142	0.4088	0.2349	0.3949	0.2807	0.2826	0.3183	0.2545	0.2111	0.1979
operprof	0.5212	0.4635	0.4014	0.1355	0.1435	0.1428	0.2650	0.1435	0.3102	0.1726	0.2594	0.2433	0.3662	0.3605	0.3861
pchgm_pchsale	0.0796	0.0605	0.0065	0.2424	0.2518	0.2691	0.4469	0.2518	0.4430	0.2571	0.2603	0.2704	0.1059	0.0774	0.1131

（续表表 C.1 各算法的因子重要性程度（详细））

ATO	0.0427	0.2015	0.1829	0.2819	0.2860	0.2869	0.5216	0.2860	0.4994	0.4028	0.4311	0.4406	0.1584	0.1598	0.1633
nincr	0.0848	0.0317	0.0569	0.1295	0.1439	0.1128	0.8031	0.1439	0.7600	0.7506	0.7332	0.7052	1.7658	1.7720	1.7028
roic	1.0247	5.2071	2.9364	0.2564	0.2630	0.3254	0.9334	0.2630	1.0597	1.8693	1.8419	2.0245	0.9538	0.9569	0.9197
rusp	0.3608	0.3030	0.3774	0.2640	0.2660	0.2453	0.4633	0.2660	0.4380	0.1162	0.1760	0.1420	0.3629	0.4039	0.3813
CR	0.6531	0.3517	0.5484	0.2225	0.2270	0.2375	0.4758	0.2270	0.4639	0.2779	0.3071	0.2982	0.2960	0.3511	0.3668
QR	0.3781	0.2746	0.4173	0.2455	0.2131	0.2311	0.4764	0.2131	0.4528	0.3158	0.3148	0.3127	0.3917	0.4152	0.4088
CFdebt	0.7686	0.8941	0.7218	0.2754	0.2240	0.2400	5.7807	0.2240	5.2061	6.8030	6.5682	6.5539	0.5281	0.5635	0.5311
salecash	0.1112	0.0430	0.1353	0.2319	0.2892	0.2677	0.5595	0.2892	0.5708	0.2898	0.3032	0.2979	0.0147	0.0689	0.0649
saleinv	0.0087	0.1046	0.0755	0.1490	0.1668	0.1740	0.0836	0.1668	0.0227	0.0908	0.1056	0.0836	0.5426	0.5239	0.5657
CRG	0.5592	0.3929	0.6733	0.2463	0.2478	0.2592	0.5160	0.2478	0.5446	0.3747	0.3960	0.4305	0.1360	0.1570	0.1519
qRG	0.5628	0.3015	0.5840	0.2653	0.2727	0.2950	0.5521	0.2727	0.5490	0.4009	0.4132	0.4405	0.2067	0.2028	0.2068
pchsaleinv	0.2613	0.1439	0.1735	0.2924	0.3292	0.3347	0.6165	0.3292	0.6219	0.4000	0.4085	0.4274	0.7572	0.7790	0.7989
salerec	0.2220	0.0878	0.1173	0.1099	0.1010	0.0887	0.0231	0.1010	0.0586	0.1164	0.0867	0.0790	0.6876	0.6821	0.7480
tang	0.0410	0.1665	0.0442	0.1724	0.1641	0.1820	0.3369	0.1641	0.3241	0.0027	0.0156	0.0584	0.3315	0.4310	0.4621
herf	0.4760	0.4763	1.1174	0.0973	0.1164	0.0605	0.6678	0.1164	0.8936	2.0576	2.2781	2.5581	1.4111	1.4842	1.4773